ANÁLISIS BIOINFORMÁTICO DE LOS EVENTOS FRACTALES RELACIONADOS CON FENOMENOS EPIGENÉTICOS DEL GENOMA HUMANO EN LA INTEGRACIÓN DE LOS LENTIVIRUS.

LINA ANDREA ALZATE RINCON

UNIVERSIDAD DEL VALLE

FACULTAD DE CIENCIAS

DEPARTAMENTO DE QUÍMICA

SANTIAGO DE CALI

2014

ANÁLISIS BIOINFORMÁTICO DE LOS EVENTOS FRACTALES RELACIONADOS CON FENOMENOS EPIGENÉTICOS DEL GENOMA HUMANO EN LA INTEGRACIÓN DE LOS LENTIVIRUS.

LINA ANDREA ALZATE RINCON

Trabajo de grado para optar por el título de químico.

Director:

Dr. FELIPE GARCÍA VALLEJO. M.S, PhD.

Codirector:

Dr. PEDRO ANTONIO MORENO. PhD.

UNIVERSIDAD DEL VALLE

FACULTAD DE CIENCIAS

DEPARTAMENTO DE QUÍMICA

SANTIAGO DE CALI

2014

UNIVERSIDAD DEL VALLE

FACULTAD DE CIENCIAS

PROGRAMA DE QUÍMICA

ALZATE RINCON, LINA ANDREA, 2014

“ANÁLISIS BIOINFORMÁTICO DE LOS EVENTOS FRACTALES RELACIONADOS CON FENOMENOS EPIGENÉTICOS DEL GENOMA HUMANO EN LA INTEGRACIÓN DE LOS LENTIVIRUS”

MATERIAS O TEMAS:

- **BIOQUIMICA**
- **BIOLOGIA MOLECULAR**
- **BIOLOGIA COMPUTACIONAL**
- **BIOINFORMATICA**
- **VIROLOGIA**

NOTA DE ACEPTACIÓN: **APROBADA**

Jurado: **Cesar Godoy PhD.**

Santiago de Cali, 15 diciembre de 2014

AGRADECIMIENTOS

Agradezco primero a Dios, que me ha dado la fortaleza para seguir adelante estos dos últimos años de investigación.

Agradezco a la Universidad del Valle y al Departamento de Química por bridarme las facilidades para el aprendizaje y los medios para llevar a término mi carrera profesional. A mi director de tesis por su guía en este proyecto, por ayudarme a formar como profesional y como persona; sus pensamientos y su confianza fueron fundamentales para orientarme los últimos años. Una mención especial, el voto de confianza que me brindó y me brinda hasta el día de hoy.

A mi codirector, por su gran contribución y guía para este trabajo. Sus consejos y pensamientos hicieron de este proyecto, un gran trabajo investigativo.

A Carlos Tellez, que es parte importante de este proyecto; gracias por su ayuda, apoyo incondicional en el desarrollo de mi tesis y por los buenos ratos. Sin tu paciencia no hubiera sido posible esto.

A mis padres, mi hermana, mis abuelos y Álvaro Loaiza; que me acompañaron y apoyaron de una u otra manera en el transcurso de este proyecto de investigación sin importar las circunstancias.

Al Laboratorio de Biología Molecular y Patogénesis y al Grupo de Bioinformática de la Escuela de Ingeniería de Sistemas; tanto profesores e integrantes.

Al departamento de Biología, la Escuela de Salud, la Escuela de ingeniería, por brindarme las herramientas académicas necesarias y las bases, para el desarrollo de este trabajo.

A mis amigos de química, biología, salud e ingeniería, por creer en mí; su apoyo fue realmente fundamental en todo este proceso que hoy culmina.

“Nunca te rindas en lo que realmente quieres hacer. La persona con grandes sueños es más poderosa que una con todos hechos.”

— Albert Einstein

TABLA DE CONTENIDO

Pág.

1. RESUMEN 11

2. INTRODUCCIÓN 13

3.1. Objetivo general 15

3.2. Objetivos específicos 15

4. MARCO TEÓRICO Y ANTECEDENTES 16

4.1. Lentivirus: características generales. 16

4.2. Síndrome de la inmunodeficiencia adquirida (SIDA): problema de salud pública. 17

4.3. Mecanismo de integración Lentiviral. 17

4.4. Estructura de la integrasa 20

4.5. Genómica del proceso de integración retroviral 21

4.6. Análisis Multifractal del genoma 23

5. METODOLOGIA 27

5.1. Universo de estudio 27

5.2. Parámetros moleculares para caracterizar las secuencias de estudio 27

5.3. Clasificación de Cromosomas por parámetros multifractales 28

5.4. Representación del juego del caos y valores de multifractalidad para las secuencias de estudio. 28

5.5. Análisis estadísticos 30

6. RESULTADOS Y DISCUSIÓN. 31

6.1. Análisis de homología de las secuencias de integración viral de estudio con respecto al genoma de referencia. 31

6.2. Características estructurales y funcionales de las secuencias asociadas a los sitios de integración. 33

6.2.1. Distribución de los elementos repetitivos de las secuencias de integración del cADN. 33

6.3. Estudio de la multifractalidad y su relación con el ambiente genómico. 35

6.3.1. Caracterización de las secuencias por valores de multifractalidad por cromosoma 36

6.3.2. Rango de valores de multifractalidad de las secuencias específicas de estudio. 37

6.3.3. Correlación entre los elementos de caracterización genómica y los valores de multifractalidad 38

6.4. Análisis de genes de las secuencias de integración de los virus VIH-1 y VIH-2 43

6.5. Análisis de rutas metabólicas para los genes de las secuencias de integración de los virus VIH-1 y VIH-2 44

7. CONCLUSIONES 45

8. PERSPECTIVAS 46

9. BIBLIOGRAFÍA 46

9. ANEXOS 52

LISTA DE FIGURAS Y ESQUEMAS

Figura 1. Representación gráfica y micrografía electrónica del virión de VIH-1. La figura muestra la localización de las proteínas que componen la partícula viral.

Figura 2. Descripción de las etapas del proceso de integración del Virus de la Inmunodeficiencia Humana tipo 1 (VIH-1) y localización de los diferentes complejos moleculares.

Figura 3. Modelación molecular de la interacción entre el dímero de la integrasa del VIH-1 con el ADN blanco de integración. Se muestra la superficie de contacto de la enzima con la doble hélice del ADN.

Figura 4. Ejemplos sobre el juego del caos. A) Representación general del CGR. B) perfil del "conteo de cajas" para una secuencia especifica del estudio de integración lentiviral en el genoma humano.

Figura 5. Ejemplos de fractales 2D y 3D. C) la figura representa un polímero o glóbulo en conformación 3D: sin nudos y en su máxima compactación. D) los pixeles negros forman un cluster en 2D que se obtiene mediante la agrupación de elementos distribuidos aleatoriamente utilizando conexiones del vecino más cercano.

Figura 6. Distribución del número de secuencias de provirus por cromosomas. A) Aspectos generales para la distribución para el VIH-1 y VIH-2 (100% de homología en los 24 cromosomas humanos). B) Por clasificación de las secuencias por cromosomas multifractales.

Figura 7. Contenido de genes, islas CpG y repeticiones evaluadas por secuencias de estudio y clasificadas y caracterizadas por cromosoma para la integración del cADN del VIH-1.

Figura 8. Contenido de genes, islas CpG y repeticiones evaluadas por secuencias de estudio y clasificadas y caracterizadas por cromosoma para la integración del cADN del VIH-2.

Figura 9. Espectro de distribución multifractal para el VIH-1 y el VIH-2, con el fin de evaluar el rango multifractal que son caracterizadas las secuencias de estudio.

Figura 10. Distribución de los valores de los promedio de multifractalidad por secuencias y clasificado por cromosoma.

Figura 11. Distribución de las secuencias de estudio del VIH-1 y VIH-2 por rango de valores de multifractalidad.

Figura 12. Valores de coeficientes de correlación de Pearson para el VIH-1 entre los valores de multifractalidad y el número de secuencias Alu e islas CpG.

Figura 13. Valores de coeficientes de correlación de Pearson para el VIH-2 entre los valores de multifractalidad y el número de secuencias Alu e islas CpG.

Figura 14. Diagrama que explica el significado de los valores de multifractalidad en el genoma humano y la tendencia a diferentes estados condicionados por la remodelación epigenética y ambiental y la determinística y genética.

Figura 15. A) Distribución en análisis de dos componentes de los cromosomas por los valores de multifractalidad. B) Análisis de los cromosomas por territorios cromosómicos der acuerdo con resultados experimentales tomados de fibroblastos humanos en fase G0 por Bolzer et al.

Figura 16. Contenido de genes de las secuencias del estudio por cromosoma para el VIH-1 y VIH-2.

Figura 17. Rutas metabólicas predominantes en las funciones de los genes que contienen las secuencias de estudio para el VIH-1 y VIH-2 en los cromosomas 16, 17 y 19. Obtenidos de KEEG (http://www.genome.jp/kegg/).

LISTA DE ABREVIATURAS, ACRÓNIMOS Y SÍMBOLOS

VIH-1 Virus de la inmunodeficiencia humana tipo I.

VIH-2 Virus de la inmunodeficiencia humana tipo II.

VIS Virus de la inmunodeficiencia en simios.

DM Dimensión multifractal.

Lentivirus virus lentos, que integran en el genoma del hospedero.

Islas CpG Regiones donde existe alta concentración de pares de citosina y guanina enlazados por fosfatos.

SINE Short Interspersed Nuclear Elements (Elementos nucleares dispersos cortos).

LINE Long Interspersed Nuclear Elements (Elementos nucleares dispersos largos).

cDNA DNA complementario.

In silico Hecho por computadora o vía simulación computacional.

NCBI National Center for Biotechnology Information.

Contig Es un conjunto de superposición de los segmentos de ADN que en conjunto representan una región de consenso de ADN.

DNA Ácido Desoxirribonucleico.

RNA Ácido Ribonucleico.

Provirus Son DNA viral integrado en el genoma de una célula huésped.

SIDA Síndrome de la inmunodeficiencia humana adquirida.

T CD4 Linfocitos T colaboradores o linfocitos T cooperadores.

OMS Organización mundial de la salud.

Gen *Pol* El gen pol codifica tres enzimas: la transcriptasa inversa (RT), una ribonucleasa (PR) y una integrasa (IN).

RNAsa Ribonucleasa.

LTR Repetición terminal larga (long terminal repeat).

Alu Secuencias altamente repetitivas reconocida por la enzima de restricción Alu I.

Hotspot Secuencias o sitios calientes a la integración.

In vitro Realizar un determinado experimento en un tubo de ensayo.

In vivo Que ocurre o tiene lugar dentro de un organismo.

CGR Representación juego del Caos.

N(E) Numero partes iguales requeridas cuando el factor E es aplicado.

D Dimensión Fractal.

Dq Algoritmo Fractal.

1. RESUMEN

La integración retroviral del VIH en el genoma humano es uno de los procesos más complejos en el estudio de la dinámica estructural y funcional del virus durante la progresión de la enfermedad. Uno de los principales problemas en el estudio de la integración es la existencia de muchos sitios calientes (“hotspot”), de preferencia viral y la razón por la que el virus prefiere estas zonas es aún desconocida. La evidencia actual muestra que la integración de cADN lentivirus no es al azar, ya que algunas condiciones topológicas de la cromatina interfásica son importantes para determinar el “nicho” genómico para integrarse. Estudios previos han demostrado que la multifractalidad a lo largo del genoma humano ésta relacionada con el contenido de elementos Alu y la estabilidad del genoma. Esto es debido que las regiones de baja y media multifractalidad, de baja estabilidad genómica, son más susceptibles de exhibir efectos epigenéticos-ambientales, mientras que las regiones de alta multifractalidad, más estables, estarían relacionadas con efectos genético-deterministas. Dado que el VIH es un agente exógeno al genoma, se predice que los sitios de integración del VIH estarían relacionados con valores de baja y media multifractalidad y por ende de baja estabilidad genómica. El objetivo de la presente investigación es poner a prueba el modelo multifractal propuesto versus la integración del VIH.

A fin de caracterizar los sitios de integración específicos para los virus VIH-1 y VIH-2, 2184 secuencias (2098 secuencias de VIH-1 y 191 de VIH-2) de 100 Kbp de longitud alrededor del sitio de integración fueron obtenidas a partir de regiones flanqueantes a extremos LTR de provirus en macrófagos y células mononucleares de sangre periférica depositadas en el NCBI. Estas secuencias fueron alineadas con el genoma humano, utilizando un algoritmo de búsqueda que arroja un porcentaje de homología con el genoma humano referencia (versión GRCh37); mediante este procedimiento, se obtuvo información sobre su localización cromosómica y posición por Contig. Posteriormente, las secuencias de los distintos alineamientos se analizaron por medio del análisis multifractal y estas, fueron

correlacionadas con las variaciones estructurales y moleculares encontradas en dichas secuencias, como contenidos de secuencias Alu, islas CpG, genes, etc. Los resultados permitieron descubrir que el VIH-1 se integra en regiones de baja y media multifractalidad, mientras que el virus VIH-2 lo hace en regiones de media multifractalidad. Igualmente, se encontraron correlaciones altas entre la multifractalidad de las secuencias y los contenidos de secuencias repetidas del tipo Alu e islas CpG para ambos tipos de virus. Con base en los resultados obtenidos, se propone un modelo no lineal descriptivo para el proceso de integración lentiviral del VIH-1 y VIH-2, con algunas implicaciones biológicas. El modelo revela que la integración lentiviral para los dos tipos de virus está localizada en regiones de baja y media estabilidad de la cromatina. Esta organización no lineal de integración apoya el modelo de predicción propuesto, lo cual tiene un significado biológico relevante para comprender el papel de la interacción entre el virus y el hospedero humano. Igualmente, demuestra que las regiones de integración están relacionadas con genes localizados en regiones de baja y media multifractalidad y bajos contenidos de secuencias Alu. Las familias de genes encontradas en el estudio principalmente están localizadas en los cromosomas 19 y 17 y estos fueron: TM4SFS, DOHH, MSFD12 y GLTP. Estas familias esenciales en codificación de proteínas de señalización celular y ciclo celular en linfocitos T.

Palabras claves: Análisis multifractal, integración de VIH, genoma humano, relación epigenética, variaciones estructurales en la integración.

2. INTRODUCCIÓN

Los virus las inmunodeficiencias humanas (VIH), son agentes asociados con el Síndrome de Inmunodeficiencia Adquirida (SIDA). Actualmente existen dos tipos de VIH (VIH-1 y VIH-2) que de acuerdo con el Comité Internacional de Taxonomía de Virus (ICTV), se incluyen en el género Lentivirus, dentro de la subfamilia Orthoretrovirinae de la familia Retroviridae. El VIH-1 fue descubierto e identificado, como el agente de la actual pandemia del SIDA, por el equipo de Luc Montagnier en Francia en 1983 [1]; posteriormente, en 1985, el grupo de Max Essex en Estados Unidos, descubrió el VIH-2 en pacientes Senegaleses con SIDA [2].

Morfológicamente la partícula del VIH mide aproximadamente100 nm de diámetro, es de forma esférica y está rodeada por una bicapa fosfolipídica de la célula hospedera que la incorpora durante el proceso de gemación. Su genoma son dos moléculas de ARN de aproximadamente 9000 nucleótidos no complementarias entre si y que contienen los genes: gag (proteínas de la matriz y de la cápside); pol (codifica las enzimas transcriptasa reversa, proteasa e integrasa); env, codifica para las proteínas de envoltura (gp120 y gp41); además de seis genes adicionales vif, vpr, rev, vpu, tat y nef necesarios para la replicación e infección [3] (Figura 1).

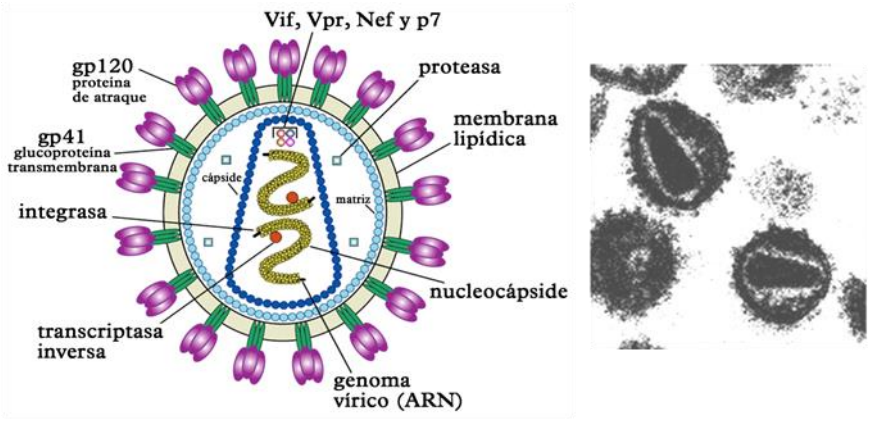

Figura 1. Representación gráfica y micrografía electrónica del virión de VIH-1. La figura muestra la localización de las proteínas que componen la partícula viral. Fuente: National Institute of Healt. [http://www.niaid.nih.gov NIAID]

Hay muchos mecanismos de latencia del VIH que podrían explicar la activación de la integración en sitios que afectan la transcripción, ya que la activación o represión de la transcripción puede ser crucial para la propagación del virus [4]. Además la complejidad de este virus se asocia con respecto al genoma del huésped; el genoma humano es una de las estructuras más complejas nunca antes vistas en la naturaleza; extraordinariamente está altamente organizada y esto son motivos de la variedad de las funciones entre las que se encuentra la regulación génica, transcripción RNA y DNA metilación etc. [5-8]. La integración proviral en sitios específicos para células T Jurkat ha revelado que, la latencia de las células infectadas con VIH está en zonas que pueden ser inactivas y que posteriormente son reactivados al momento de la transcripción. En forma general los productos de la latencia de los lentivirus como los estudiados, producen silenciamiento de ciertos genes de importancia celular que pueden facilitar la infección o proliferación del virus [9].

Existen múltiples sitios de integración que han sido secuenciados y mapeados en relación con las características de la cromatina tales como las unidades de transcripción, islas CpG y promotores; en general los retrovirus tienen distintos sitios de integración como por ejemplo el VIH que integra preferencialmente en unidades de transcripción [a]. Varios estudios previos *in silico* han descrito ciertos perfiles de integración simultáneos del VIH-1 y VIH-2; prediciendo con esto que ambos lentivirus muestran tendencia a la integración en genes que codifican proteínas es decir, en regiones altamente génicas [10].

La no linealidad de los procesos biológicos, es una propuesta novedosa para explicar cuál es el papel de los factores que intervienen al momento de la integración del virus y la preferencia del retrovirus para la integración en el genoma humano. Además esto se ha relacionado con cambios estructurales aplicando teoría del caos empleando la simulación fractal a la estructura y dinámica del genoma humano [11].

Este trabajo considera reportes en la base de datos de algunos posibles sitios de integración de los lentivirus VIH 1 y 2, para ser evaluados no solo por sus características moleculares, considerando secuencias repetitivas como SINEs y LINEs a la vez que islas CpG, además de relacionar la fractalidad y estructura del genoma humano como otra variable no lineal que permita explicar las de los perfiles de integración de los lentivirus en el genoma humano. Para su ejecución se realizó *in silico* a través de análisis bioinformático aplicando del formalismo multifractal y juego del caos. Para tal fin, se desarrollaron diferentes algoritmos de búsqueda y estudio del genoma humano con respecto a secuencias específicas reportadas en el NCBI donde integra el virus VIH en el genoma humano.

Los resultados proponen que existe ambiente genómico de un complejo estructural de las secuencias que son bancos de integración de los lentivirus, este ambiente pudo ser analizado por el fenómeno de la multifractalidad describiendo de forma definida los factores que influyen en todo el proceso de interacción entre huésped humano y virus. En este estudio, se encontró una alta correlación entre la multifractalidad de las zonas específicas con la integración proviral.

3. OBJETIVOS

3.1. Objetivo general

Con el fin de determinar la no linealidad del genoma en relación al comportamiento viral, al momento de la integración de los lentivirus VIH-1 y VIH-2, se estudiará el grado de multifractalidad del genoma humano, en aquellas zonas en las que se registra el proceso de integración del cDNA Lentiviral;

3.2. Objetivos específicos

- Obtener información de NCBI sobre las secuencias que flanquean sitios de integración en una extensión de ±100Kpb.

- Construir una base de datos que contenga información sobre Contig, la posición de la secuencia en el genoma y genes contenidos en linfocitos TCD4+.
- Evaluar por multifractalidad las secuencias que flanquean los sitios de integración de los diferentes lentivirus a partir de las bases de datos construida y sus características correspondientes.
- Calcular la probabilidad de que las secuencias escogidas sean o no fractales en la integración.
- Estudiar las funciones de los distintos genes clase II en donde se registra altos índices de integración y correlación con los valores de multifractalidad.

4. MARCO TEÓRICO Y ANTECEDENTES

4.1. Lentivirus: características generales.

Históricamente, los lentivirus han sido investigados a lo largo de los años. El diverso grupo de virus que compone los lentivirus tiene características comunes y distintivas entre sí [12]. El ciclo de vida y en particular como el ciclo celular del hospedero influye en la replicación lentiviral, ha sido muy discutido. Desde esta perspectiva, los eventos de replicación lentiviral han sido caracterizados en primates y humanos: VIH-1, VIH-2 y VIS [13]; por eso el estudio exhaustivo de esta clase de virus es clave para el desarrollo y tratamiento del SIDA.

Los lentivirus son clasificados en una gran familia de Retroviridae, estos han sido definidos como virus de RNA que en forma obligatoria se propagan a través de un intermediario intracelular constituido por una molécula de DNA de cadena doble sintetizada por la enzima transcriptasa reversa, característica de estos virus, a partir del RNA que constituye el genoma viral; sin embargo para que sean exitosos en su

ciclo de vida, se deben integrar al genoma de la célula blanco y producir un estado proviral estable [14].

4.2. Síndrome de la inmunodeficiencia adquirida (SIDA): problema de salud pública.

El síndrome de la inmunodeficiencia humana adquirida (SIDA) fue descrito por primera vez a mediados de 1981, el agente etiológico es un retrovirus que se conoce con el nombre de virus de la inmunodeficiencia humana (VIH). No se conoce enfermedad reciente en la historia de la humanidad que haya mantenido tan ocupados a la comunidad científica y a la población en general en la búsqueda de una solución definitiva [15].

En general se ha observado que este virus tiene efectos severos en el sistema inmune; una importante pista sobre esto es que en pacientes con SIDA con las infecciones pulmonares causadas por un hongo llamado *Pneumocystis carini.* Esta infección es muy rara en individuos sanos, pero en pacientes con cáncer del sistema inmune (linfoma) es muy común. En gran parte de los pacientes que tienen SIDA confirmado tienen un daño inminente en el sistema inmune [16].

4.3. Mecanismo de integración Lentiviral.

La integración del ADN retroviral en el genoma del hospedero es un paso esencial para la replicación del virus en la célula infectada. Es así como durante las primeras etapas de la infección que preceden a la integración, el ARN viral junto con la Transcriptasa Reversa (TR), la integrasa (IN) y la proteína de la matriz (MA) [17-19] reclutan, en el citoplasma, una serie de proteínas celulares para conformar el complejo de preintegración (CPI). En el transporte del CPI del citoplasma al núcleo, la proteína viral Vif parece actuar como puente de unión entre el CPI y los microtúbulos promoviendo su transporte activo hacia el núcleo [20].

A diferencia de los Oncorretrovirus como los HTLV, los Lentivirus humanos poseen una baja dependencia del ciclo celular y pueden infectar células quiescentes. Esto es posible ya que mediante un proceso de translocación, el CPI atraviesa la envoltura nuclear; en este paso intervienen tanto la Vpr [10] como las proteínas MA e IN [11-13] que forman un complejo proteico heterodimérico compuesto además por Importina-a (Imp-a) e Importina-b (Imp-b). Mientras que la Imp-a se une a secuencias específicas NLS (Señales de Localización Nuclear), la Imp-b participa en la translocación del CPI a través del complejo del poro nuclear [14]. Se ha determinado que tanto la proteína MA como la IN, contienen secuencias NLS que pueden interactuar con miembros de las importinas [18].

La reacción de la integración se produce en tres etapas, en la primera etapa, el procesamiento del extremo 3', recorta los primeros nucleótidos de cada extremo 3' del ADN viral, dejando un grupo -OH de un dinucleótido CA altamente conservadas [22], este escisión tiene lugar en el citoplasma. El segundo paso, el extremo 3' de unión, tiene lugar en el núcleo, aquí se cataliza el ataque del grupo CA-OH en el ADN viral en el ADN celular, esta es una reacción de transesterificación de un solo paso este ataque concertado produce dos brechas en cada hebra del ADN del huésped, y como consecuencia de la reparación de estas lagunas, que se cree que puede hacer por las proteínas del huésped, dos secuencias similares del flanco ADN celular del provirus integrado (Figura 2).

Se ha propuesto que la integración retroviral no es al azar y que, por el contrario, existen regiones del genoma en donde habría una de alta probabilidad de integración del cADN viral [16-19]. Sin embargo, a pesar de la evidencia actual, el mecanismo por el cual se seleccionan estas regiones de mayor probabilidad, todavía no se conoce completamente. Ésto es porque la integración es un proceso multifactorial el cual varía entre las especies de retrovirus y está influenciado principalmente por las variaciones epigenéticas intrínsecas de la cromatina

hospedera, además de las proteínas celulares y virales que se unen al sitio blanco [20, 22].

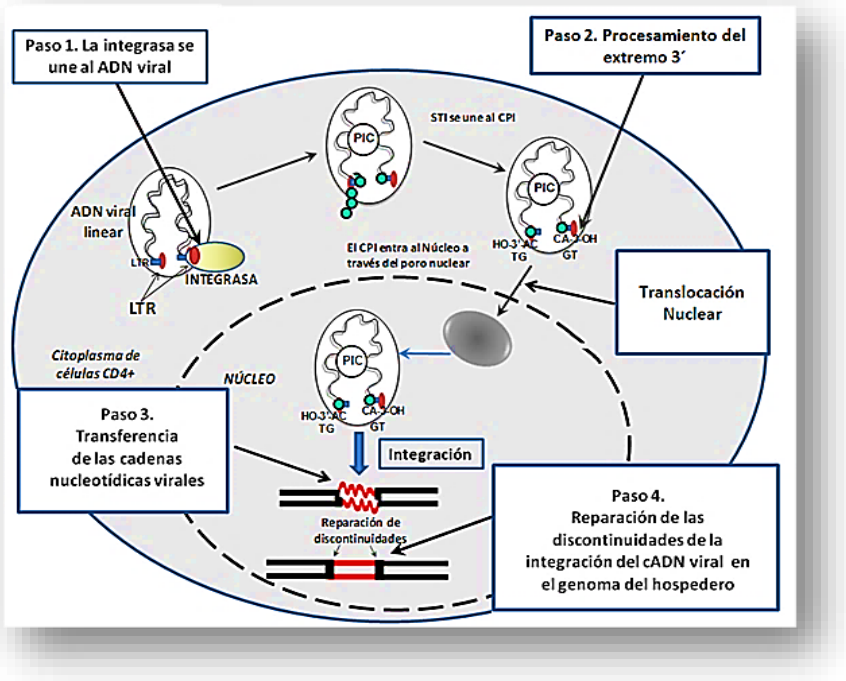

Figura 2. Descripción de las etapas del proceso de integración del Virus de la Inmunodeficiencia Humana tipo 1 (VIH-1) y localización de los diferentes complejos moleculares. Fuente: Laboratorio de Biología Molecular y Patogénesis, LABIOMOL. Universidad del Valle.

Estudios iniciales sobre la integración retroviral han demostrado que la distribución de provirus en el genoma humano no es al azar y que existen zonas consenso del genoma hospedero que presentan características comunes [22]. Éstas, de acuerdo con sus características estructurales y de remodelación de la arquitectura cromatínica por mecanismos epigenéticos, definen un ambiente genómico especial que facilita una mayor frecuencia de integración del ADN viral [20]. Así pues, el ambiente genómico es una zona de la cromatina con características diferenciales que incluyen secuencias repetitivas, islas CpG, contenido de Guanina-Citosina, densidad de genes, entre otras, que se utilizan entre otros, para estudiar a nivel

genómico procesos biológicos epigenéticos que ocurren como resultado de la integración retroviral.

En investigaciones recientes se ha demostrado que la alta organización estructural y el entendimiento de secuencias repetitivas del DNA muestran un nuevo escenario para el estudio biológico del proceso de integración del cDNA de los lentivirus. Hay muchas regiones preferenciales de sitios de integración: como lo son las secuencias repetitivas que están a largo del genoma; en muchos estudios *in vitro* e *in vivo* muestran que la integración del VIH-1 predomina activamente en la trascripción donde hay zonas de alta densidad génica, frecuencias de Alu y alto contenido de islas CpG [23]. Se le denominan a estos sitios de integración Hotspot (alta frecuencia de eventos de integración).

4.4. Estructura de la integrasa

La integrasa retroviral, se puede subdividir en tres dominios sobre la estructura y el comportamiento, estos dominios son, el dominio amino terminal (-NH2), caracterizado por una estructura de dedo de zinc, llamado el dominio HHCC por la conservación de dos histidina y dos cisteínas, entre los que el número de residuos es variable; mutaciones en esta región pueden afectar la interacción de la enzima con ADN viral. El segundo dominio es el núcleo catalítico, que se encuentra aproximadamente entre los residuos 50 y 212, llamado DD-35-E a causa de los aminoácidos invariantes aspártico (D) y glutámico (E) (el 35 se refiere a los residuos entre ellos), las mutaciones en esta región podrían suprimir todas las reacciones realizadas por la enzima. El dominio carboxílico (-COOH), es el dominio con menos conservación de aminoácidos entre los retrovirus; puede ser esencial para la interacción con otros componentes de la maquinaria de integración [24].

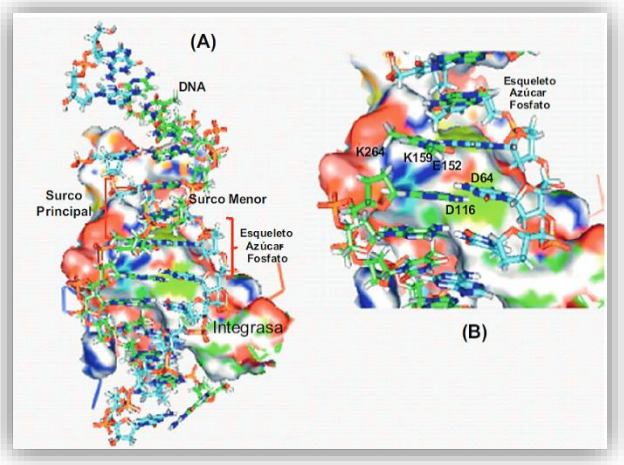

Figura 3. Modelación molecular de la interacción entre el dímero de la integrasa del VIH-1 con el ADN blanco de integración. Se muestra la superficie de contacto de la enzima con la doble hélice del ADN. A) Vista panorámica de la interacción en la que se precisan los dominios del sitio activo, compuesto, entre otros, por los residuos catalíticos D64, D116 y E152; además de la región de interacción con el ADN con el K264. B) Aproximación ampliada que muestra la interacción entre el K264 y una porción del esqueleto azúcar fosfato del ADN blanco. En el interior de la integrasa se localizan los aminoácidos catalíticos, interactuando con la porción interna de una vuelta del ADN blanco. Las figuras se obtuvieron a partir de la estructura de la integrasa del VIH-1 código PDB 1BIS, que fue obtenida por difracción de rayos X. Para ello se utilizó el programa PyMol. (D) ácido aspártico; (E) ácido glutámico; (K) lisina.

4.5. Genómica del proceso de integración retroviral

La replicación retroviral requiere la integración covalente del cADN, sintetizado durante la transcripción inversa, en la cromatina de la célula huésped. La forma integrada del virus, denominado el provirus, proporciona una plantilla para la expresión génica viral. Debido a que el provirus es una parte integral del genoma del huésped, estos persisten en el huésped durante toda la vida en la célula

infectada. Este rasgo de la integración irreversible hace que los retrovirus sean vehículos especialmente atractivos para terapia genética en humanos [25]. Se ha propuesto que la integración retroviral no es al azar y que, por el contrario, existen regiones del genoma en donde habría una alta probabilidad de integración del ADN viral. Sin embargo, a pesar del conocimiento actual, el mecanismo por el cual se seleccionan estas regiones no se conoce completamente [26-28]. La integración lentiviral se considera un proceso multifactorial, en el cual varía entre especies de retrovirus y está influenciado principalmente por lo proceso epigenéticos intrínsecos de la cromatina del huésped. Un factor determinante en la integración es la arquitectura de la cromatina en ciertos estadios de remodelación, estos pueden ser o no susceptibles a la integración lentiviral depende de todo el ambiente de interacción genómico [29].

La evaluación de los sitios más susceptibles a la integración son regiones transcripcionales y se activan no solo en regiones de exones o intrones sino también alrededor de promotores para este [9]. El gran avance en la genómica de integración se ha dado gracias a la aproximación investigativa *in silico.* De acuerdo con la definición del National Center for Biotechnology Information (NCBI) la Bioinformática es la disciplina científica que combina biología, computación y tecnologías de la información. El objetivo de esta disciplina es investigar y desarrollar herramientas útiles para llegar a entender el flujo de información. Inicialmente, la bioinformática se ocupaba sobre todo de la creación de bases de datos de información biológica, especialmente secuencias, y del desarrollo de herramientas para la utilización y análisis de los datos contenidos en esas bases de datos. Sin embargo, la Bioinformática ha ido evolucionando para ocuparse cada vez con mayor profundidad del análisis e interpretación de los distintos tipos de datos (secuencias de genomas, proteomas, dominios y estructuras de proteínas, etc). Como resultado del avance de esta disciplina, la investigación biomédica moderna se apoya, por un lado, en procesos experimentales y por otro en búsqueda asistida por computador e internet de bancos de datos que guardan los datos estructurales

y funcionales de genes, proteínas y otras biomoléculas, lo que se conoce como el análisis *in silico* [30].

4.6. Análisis Multifractal del genoma

La multifractalidad es una propiedad de autosemejanza, es una característica de la geometría fractal observada en la naturaleza en general y en particular para estudios biológicos. Se ha demostrado que el mantenimiento del estado celular en un estado estacionario no garantiza las funciones mismas de la célula; en este aspecto, el comportamiento fractal se refiere a una señal biológica que conlleva cambios en la morfología en un sistema no lineal. Se predice así que el paisaje genómico tiene geometría fractal y el estudio se ha dado en parte a la búsqueda de propiedades estructurales auto-similares que intervengan en un proceso biológico de regulación celular [10] (figura 4).

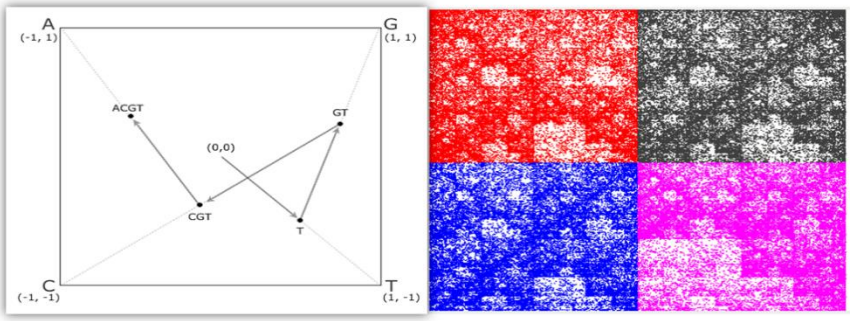

Figura 4. Ejemplos sobre el juego del caos. A) Representación general del CGR. B) perfil del “conteo de cajas” para una secuencia especifica del estudio de integración lentiviral en el genoma humano.

Los genomas han revelado una estructura compleja que es altamente regionalizado, y que puede ser estudiado por métodos que permiten medir cómo se fragmenta el contenido de la información. La teoría de la información ha sido un marco conceptual muy útil para estudiar el contenido de información a lo largo de una secuencia de símbolos como señales [31,32]. Durante los años 60’ se estableció la geometría

fractal, como una nueva geometría para medir la irregularidad de la naturaleza. El paradigma de la geometría fractal introduce varias maneras a medir este contenido de información mediante el cálculo de la dimensión fractal, un exponente que se deriva a partir de una ley de potencia, lo que nos da una idea del nivel o la información de fragmentación contenido de un fenómeno complejo [33]. La geometría fractal es un enfoque útil para la búsqueda de propiedades auto-similares en las estructuras y procesos; ha sido un enfoque útil para hacer frente a varios problemas relacionados a la codificación y secuencias de ADN no codificantes, relaciones filogenéticas, y para la búsqueda explicaciones de regularidades observadas en bases de datos moleculares. Este formalismo se aplica cuando muchos subconjuntos fractales con diferentes propiedades de escala (con un gran número de exponentes o dimensiones fractales) coexisten simultáneamente. Como resultado, cuando un espectro de singularidades de medidas multifractales se genera, el comportamiento de la escala de las frecuencias de símbolos de una secuencia se puede cuantificar [32].

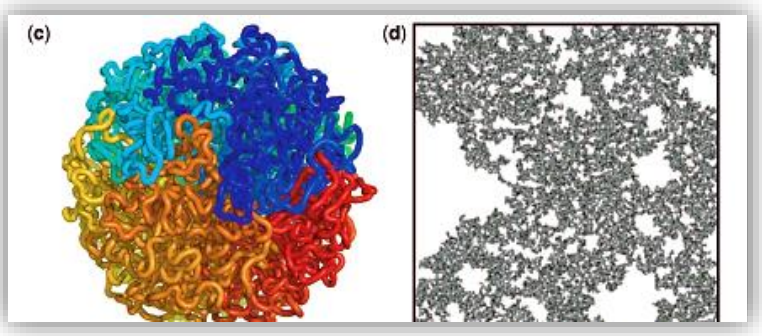

Figura 5. Ejemplos de fractales 2D y 3D. C) la figura representa un polímero o glóbulo en conformación 3D: sin nudos y en su máxima compactación. D) los pixeles negros forman un cluster en 2D que se obtiene mediante la agrupación de elementos distribuidos aleatoriamente utilizando conexiones del vecino más cercano.

El análisis Multifractal se ha implementado para mejorar la caracterización de la falta de homogeneidad espacial de ambos patrones fractales teóricos y experimentales. Por ejemplo, se ha aplicado para estudiar el fenómeno de turbulencia, de series de tiempo análisis y modelos financieros. También ha sido útil en el estudio de diferentes tipos de problemas al ADN y secuencias de proteínas, bajo dos modalidades: primero, mediante el uso de sub intervalos en un espacio unidimensional 1-D, como el espacio para representar sub series y segundo, por medio del uso de un espacio de 2-D representado el juego del caos en contexto (figura 5). La primera modalidad se aplicó en los tiempos pregenómicos, para estudiar las secuencias de ADN haciendo análisis espectrales y multifractales de mediciones. Posteriormente, en la época postgenómica, se utilizó el análisis multifractal, para discriminar de los genomas completos de bacterias y para distinguir la codificación y secuencias en las secuencias de ADN no codificante. Ha sido útil también en el análisis de bacterias para la construcción de árboles filogenéticos y para la agrupación de estructuras de proteínas [31,34].

La multifractalidad caracterizó también el genoma humano, ya que es uno de los complejos moleculares más estudiados jamás vistos en la naturaleza. Su extraordinario contenido de información ha revelado una sorprendente codificación de secuencias específicas no codificantes [35, 36]. Está altamente regionalizado e introduce patrones complejos para la comprensión de la estructura génica y repetitiva. Además de la composición de las secuencias de ADN y su papel en humanos el desarrollo, la fisiología, la medicina y la filogenia. La secuenciación del genoma humano reveló un número controvertido de genes interrumpidos o genes clase II (25000 - 32000) con su reglamentación secuencias [35, 37] que representan alrededor del 2% del genoma. Estos genes se encuentran inmersos en un mar gigante de diferente tipos de secuencias no codificantes que comprenden alrededor 98% del genoma humano. Las regiones no codificantes se caracterizan por muchos tipos de secuencias de ADN repetitivas, donde casi 10,6% del genoma humano consiste en secuencias Alu, un tipo de SINE (elementos cortos intercalados) [38].

Estos elementos no se distribuyen al por todo el genoma, sino más bien están sesgados hacia regiones ricas en genes [39]. Pueden actuar como inserciones y la gran mayoría parece ser genéticamente inerte [40]. LINES, MIR, MER, LTR, de ADN e intrones son otros tipos de secuencias no codificantes, que representan alrededor del 86% del genoma. La nueva era de los métodos de secuenciación masiva ha permitido secuenciar más de mil genomas de humanos [35, 40] mostrando la variación genética entre los diferentes grupos humanos. En anteriores estudios se estudió el genoma humano con el método de la multifractalidad descubriendo que todas estas características anteriormente mencionadas hacen que el estudio por esta metodología caracterice y explique ciertas zonas o sectores y su función.

La metodología derivada de la geometría fractal es un enfoque muy útil para estudiar el grado de fragmentación (o irregularidad) natural, artificial y estructural o estadística [35, 38]. Las estructuras fractales se caracterizan por la auto-similitud, una dimensión fractal, la ley del exponente [40]. Sin embargo, debido a la complejidad del genoma humano, un exponente puede no ser suficiente para caracterizar un fenómeno complejo. Para esta investigación el formalismo multifractal nos puede permitir el uso de más exponentes, en este caso, el objeto de análisis es dividido en varios conjuntos fractales, cada generación de una dimensión fractal que luego se traduce en un continuo espectro de exponentes. El grado multifractalidad (DM) obtenido a partir este espectro continuo permite la medición de la información genética contenido en un genoma. Los sistemas multifractales son comunes en la naturaleza, especialmente en geofísica.

La propuesta de este trabajo, es construir características específicas del fenómeno de integración lentiviral en el genoma humano con el fin de usar la herramienta de multifractalidad como una aproximación dinámica del ambiente genómico transcripcional y post traduccional de la cromatina interfásica asociada a la integración de los Lentivirus humanos [41-42].

5. METODOLOGIA

Se llevó a cabo un completo análisis bioinformático a partir del formalismo multifractal de secuencias específicas de los lentivirus humanos; para ello se usaron de variables genómicas, tales como, el contenido de secuencias repetitivas, alta densidad génica e islas CpG, con el fin de evaluar el ambiente genómico asociado a la integracion lentiviral..

5.1. Universo de estudio

Se escogieron 2184 secuencias del genoma humano, las cuales eran 2098 para el VIH-1 y 191 para el VIH-2; flanqueantes a extremos LTR de provirus en macrófagos y células mononucleares de sangre periférica depositadas en Nucleotide de National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/). Estas secuencias provenían de pacientes infectados por el VIH, excluyendo la de los vectores lentivirales construidos.

Las secuencias se sometieron a una evaluación de homología con respecto al Hs-refseq genoma de referencia que fue descargado del sitio web de NCBI; los primeros resultados obtenidos fueron a partir de un primer filtro que ajusto las condiciones a un 100% homología para ambos lentivirus con respecto al genoma de referencia. Debido al bajo número de secuencias de VIH1 obtenidas a 100% de homología, se amplió la ventana de estudio a valores de homología entre 80 y 100% con lo que se incrementó el número de secuencias para ser analizadas.

5.2. Parámetros moleculares para caracterizar las secuencias de estudio

El contenido C+G fue contado para cada fragmento de ADN de 200 kb por un script escrito en Python. Asimismo, se incluyeron varios parámetros moleculares obtenidos de diferentes bases de datos: CGI, seq_cpg_islands.gz, Alu, LINES, MIR, MER y LTR del archivo seq_repeat.md.gz, los genes del archivo seq_gene.md.gz, exones e intrones del archivo gbk.gz y el número de funciones de genes del archivo rna.q.md.gz. Todos estos archivos fueron descargados del sitio web de NCBI

(ftp://ftp.ncbi.nlm.nih.gov/). Para evaluar funciones y rutas metabólicas de los genes que flanquean en estos específicos sitios se utilizó la base de datos KEGG (ftp://ftp.genome.jp/pub/kegg/pathway/ organismos/hsa/).

5.3. Clasificación de Cromosomas por parámetros multifractales

Se analizó la integración diferencial para ambos lentivirus en cromosomas con diferentes grados de multifractalidad de acuerdo con la clasificación propuesta por Moreno et al [11]. En el grupo A los cromosomas de baja multifractalidad (4, X, 13, 5, 18, 36, Y, 8, 2, 11); para el grupo B cromosomas de media multifractalidad (21, 14, 9, 10, 7, 12, 1, 20, 15) y por último el grupo C para cromosomas de alta multifractalidad (16, 22, 17, 19).

5.4. Representación del juego del caos y valores de multifractalidad para las secuencias de estudio.

Los parámetros moleculares evaluados, fueron comparados con perfiles reportados en la literatura como G+C y Alu; en general una geometría fractal está dada en una dimensión fractal de la siguiente manera:

$N(E) \infty E^{-D}$ (Ec. 1)

Donde N(E) es el numero partes iguales requeridas cuando el factor E es aplicado.

La dimensión fractal obtenida como

$D = \ln(N(E))/\ln(E)$(Ec. 2)

La dimensión fractal es obtenida por el “conteo de cajas” o más conocido como “box-counting” que son algoritmos que permiten convertir una figura en distintas cajas de diferentes tamaños determinados $\varepsilon = 1/E$. El análisis multifractal es usado cuando hay reglas de múltiples escalas aplicadas, así que en este caso no habría un solo espectro fractal de dimensión *Dq* para la integración de todas las *q*.

El algoritmo obtenido en el análisis de fractalidad del genoma en general es:

$$D_q = \frac{\ln\left(\sum_i \frac{M_i}{M_0}\right)^q}{\ln(\varepsilon)} \frac{1}{q-1}$$ (Ec. 3)

Donde M_i es el número de puntos en la caja con relación al número total M_0 y el tamaño de la caja. El espectro multifractal es obtenido como el límite:

$$D_q = \lim_{\varepsilon \to \infty} D_q\,(\varepsilon)$$ (Ec. 4)

Un alto valor de Dq destaca para la riqueza en la estructura y propiedades en estas regiones. Valores negativos de Q relacionan regiones dispersas; un alto Dq indica una estructura compleja. En las aplicaciones del mundo real, el límite Dq es fácilmente aproximado, a partir de los datos por un ajuste lineal [11].

$$ln\left(M_i^q\right) = D_q(\varepsilon)(q-1)\ln(\varepsilon) + (q+1)\ln(M_0^q)$$ (Ec. 5)

En el CGR del estudio determinado se traza en un cuadrado, con cada uno de los cuatro vértices etiquetados como las cuatro bases de nucleótidos A, T, G y C, respectivamente. Para inicializar, colocamos el primer punto en el centro de la plaza. El segundo punto se coloca como un punto medio entre el punto inicial y las coordenadas del vértice correspondiente a los primeros nucleótidos de la secuencia de ADN. El siguiente punto, corresponde al segundo nucleótido, se coloca como un punto medio entre el punto y previamente trazada de coordenadas del vértice. El proceso se repite para la secuencia completa y todo el genoma se trazan en cuadrado de dos dimensiones. La frecuencia de palabra diferente longitudes puede ser extraída mediante la división del espacio de CGR con una rejilla de tamaño apropiado.

Para este trabajo fue construida una caja de (512x512) para mapear las secuencias específicas de estudio de integración lentiviral. Los cuatro nucleótidos fueron asignados en los vértices de CGR como A (0,0); T (512,0); G (512,512); y C (0, 512). Si bien el cálculo para cada uno de los cuadrantes correspondientes a un nucleótido se divide en cuatro partes. El primer nucleótido de cada parte es entonces

etiquetado como para el cuadrante original de CGR, mientras que el segundo nucleótido se etiqueta como el cuadrante siguiente.

Para el análisis multifractal de los “hotspot” de la integración proviral propuesta en este trabajo, se trabajó con CGR para cada una de las secuencias y se le asignaron valores en el espectro multifractal de 20 a -20. Para caracterizar cada secuencia con respecto a los cromosomas en donde pertenece el sitio de integracion viral se promediaron estos valores. Todo esto también se evaluo en un espacio de 200.000 pares de bases (200Kpb) y se estudiaron dos zonas específicas denominadas corriente arriba (Upstream, +100Kpb) y corriente abajo (Downstream, -100Kpb).

5.5. Análisis estadísticos

Se realizó un análisis de regresión simple, determinando algunos parámetros moleculares con relación al comportamiento de los valores multifractales [10].

El conjunto de datos de las secuencias y cada conjunto de fragmentos de cromosomas fueron analizados por regresiones simple y multivariante, utilizando “R Project for Statistical Computing”, para determinar la bondad del ajuste de varios parámetros moleculares frente a los valores de multifractalidad.

Para clasificar a los cromosomas humanos con base en su multifractalidad, se generó un análisis de agrupación mediante el uso de la agrupación jerárquica empleando el programa Explorer versión 3.5 (HCE3.5). La agrupación de árboles fue generada usando los siguientes parámetros: fila por fila normalización por el control; el método de vinculación completa y de coeficiente de correlación. A partir de estos datos se generaron las correlaciones entre diferentes cromosomas y su respectiva clasificación de acuerdo con los valores de multifractalidad de las secuencias de estudio.

Según lo propuesto anteriormente se evaluaron las características de las secuencias Alu, las islas CpG, las SINES y las LINES. La evaluación de la multifractalidad se obtuvo aplicando el algoritmo multifractal desarrollado en

lenguaje Python. Estos resultados se reportaron para secuencias de integración de los virus VIH-1 y VIH-2.

6. RESULTADOS Y DISCUSIÓN.

6.1. Análisis de homología de las secuencias de integración viral de estudio con respecto al genoma de referencia.

De las 2098 secuencias reportadas en el NCBI para el VIH-1 solamente fueron analizadas 1956 pues contenían toda la información necesaria para realizar el análisis multifractal. De otra parte para el VIH-2 de 191 reportadas se analizaron 176 que cumplieron os requisitos del estudio. Se obtuvieron 115 “hits” con homología 100% para el VIH-1, 1949 “hits” con homología entre 80-99% para el VIH-1, y 116 “hits” con homología 100% para el VIH-2. Para el VIH-1 exceptuando el cromosoma Y y el 21, en el resto de cromosomas se registraron integraciones en número variable. Para el VIH-2, no se presentan integraciones en los cromosomas 4, 13, 20, 22 y Y.

Para la clasificación de cromosomas por multifractalidad se recurrió a la clasificación propuesta en el trabajo del genoma multifractal anteriormente descrita en la metodología. Cuando se simuló el VIH-1 a 100%, el 40% se localizó en los cromosomas del grupos A (46/115), el 40,8% en los cromosomas del grupo B (47/115) y el 19,13% los del grupo C (22/115) (figura 6).

Con respecto a la selección anterior de las secuencias que tenían homología entre el 80% y 100%, se amplió el panorama de integración viral para abarcar más lugares del genoma y predecir el comportamiento de interacción entre el VIH-1 y el genoma humano con una muestra más amplia; en esta, el 27,1% se localizó en los cromosomas del grupos A (527/1945), el 63,23% en los del grupo B (1230/1945) y el 9% en los del grupo C (188/1945).

En la evaluación las secuencias del VIH-2 para 100% de homología se determinó que el 37,93% se localizó en los cromosomas del grupo A (44/116), el 31,03% en los del grupo B (36/116) y el 31,03% en los del grupo C (36/116). A diferencia del VIH-1 en el VIH-2 la mayor frecuencia de integraciones ocurrió los cromosomas 6, 17 y 19 (figura 6).

Los resultados obtenidos en este trabajo muestran que, el aumento de secuencias homologas entre 80-100 puede ser explicado con base en el grado de homología variable de los genomas humanos, el cual se determina por los polimorfismos de un solo nucleótido SNP. Los estudios de haplotipificación SNP se basan en una evaluación inicial de variación de nucleótidos para identificar sitios en la secuencia de ADN que albergan variación entre los individuos. El estudio de la variación SNP ha permitido hacer una diferenciación genómica de los individuos de la especie humana [43]. Como la variación SNP cambia de un individuo a otro, el bajo número de secuencias homologas al 100% se puede explicar con base en esta variación SNP de los genomas humanos, pues las secuencias adyacentes a la integración del VIH-1 provienen de la secuenciación de diferentes individuos reflejando el alto grado de polimorfismo SNP en los sitios de integración del VIH-1.

Por otro lado el hecho de que muchas de las secuencias del VIH-2 tuviesen homología del 100% con respecto al genoma de referencia usado a diferencia del VIH-1; puede explicarse debido a los procesos de expansión clonal y retro transposición que para este caso en específico son más limitados en le VIH-2 en comparación del comportamiento de estas dos variables en el VIH-1,que a su vez está relacionado con la severidad que tiene como característica el VIH-1 en el progreso de la enfermedad.

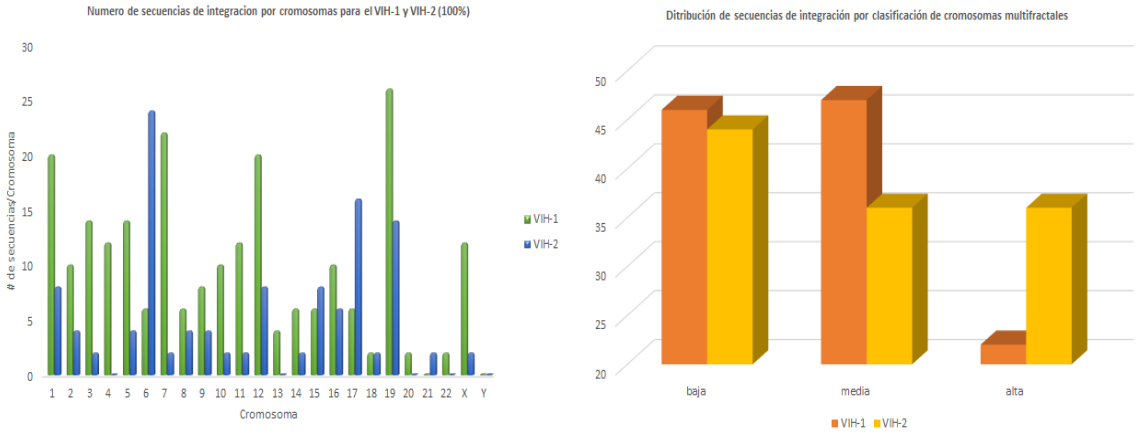

Figura 6. Distribución del número de secuencias de provirus por cromosomas. A) Aspectos generales para la distribución para el VIH-1 y VIH-2 (100% de homología en los 24 cromosomas humanos). B) Por clasificación de las secuencias por cromosomas multifractales.

6.2. Características estructurales y funcionales de las secuencias asociadas a los sitios de integración.

6.2.1. Distribución de los elementos repetitivos de las secuencias de integración del cADN.

En su conjunto, se observaron diferencias estadísticamente significativas con el contenido de repeticiones por cromosoma para el VIH-1. Se determinó que en los cromosomas 19, 16 y 17, en donde se presentaban más eventos de integración; además de los cromosomas anteriores en el VIH-2 se registraron integraciones en el cromosoma 6. Estos cromosomas son los que presentaron mayor cantidad en repeticiones como Alu, SINES, islas CpG y densidad de genes. (Figura 7).

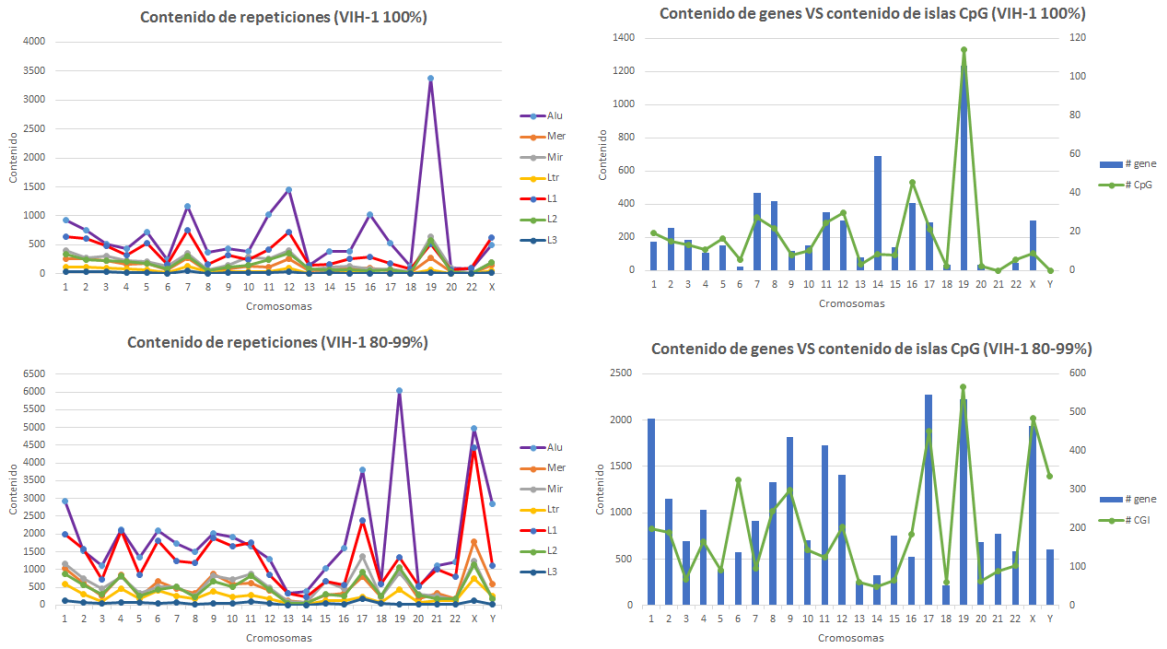

Figura 7. Contenido de genes, islas CpG y repeticiones evaluadas por secuencias de estudio y clasificadas y caracterizadas por cromosoma para la integración del cADN del VIH-1.

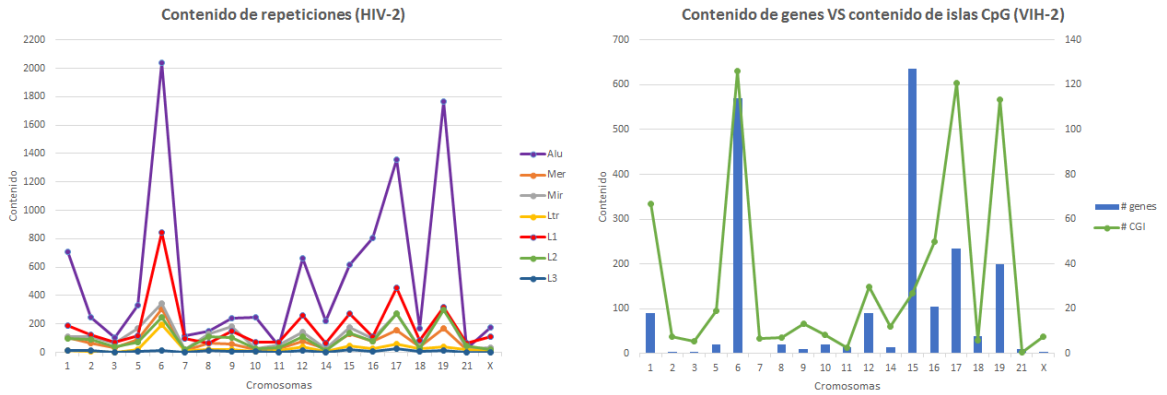

Figura 8. Contenido de genes, islas CpG y repeticiones evaluadas por secuencias de estudio y clasificadas y caracterizadas por cromosoma para la integración del cADN del VIH-2.

Resultados previos con relación a las características de los sitios de integración, han permitido proponer la influencia diferencial de factores que definen la

conformación de las zonas del genoma humano con una elevada frecuencia de integraciones lentivirales. Los perfiles de integración del virus de la leucemia murina (MLV) apoyan los resultados obtenidos en este trabajo al identificarse regiones de la cromatina abierta en unidades de transcripción. Características asociadas, tales como sitios de DNasa I-hipersensibles [44, 45] o islas CpG, estaban aparentemente enriquecidos cerca de sitios de integración del MVL. Para el virus de la inmunodeficiencia humana tipo 1 (VIH-1), se ha propuesto que la integración puede ser favorecida cerca de elementos repetitivos (incluyendo elementos LINE-1 o islas Alu [46]) o en los sitios de escisión de la topoisomerasa [ap]. En nuestro estudio, las repeticiones más numerosas en los sitios de integración fueron los Alu y los LINE-1, tanto para el VIH-1 y VIH-2. En su conjunto los resultados de éste y otros trabajos, proveen una fuerte evidencia de que ciertas áreas de la cromatina celular son más susceptibles que otras existiendo una dependencia de ciertas variables genómicas que condicionan el ambiente genómico de integración.

6.3. Estudio de la multifractalidad y su relación con el ambiente genómico.

La multifractalidad fue la herramienta utilizada para el estudio de secuencias de integración viral, este estudio se basó en el hecho que la cromatina tiene cambios conformacionales, estos cambios conformacionales permiten inferir y predecir el entorno genómico propicio de integración. Uno de los objetivos de este estudio fue evaluar con base en los valores de multifractalidad de las secuencias del genoma humano vecinas a sitios de integración de cADN lentiviral, una variable que permite valorar la dinámica de la cromatina y sus cambios estructurales que pudiesen influenciar la integración lentiviral.

Para el desarrollo del estudio de los valores de multifractalidad se verificó el espectro de las secuencias del genoma humano y el rango que caracteriza la integración proviral en las secuencias. Para el VIH-1, se obtuvo el valor mínimo ΔDq de 0,60 y valor máximo de 1,22; mientras que para el VIH-2 un valor mínimo de 0,63 y un valor máximo de 1,21 (figura 9).

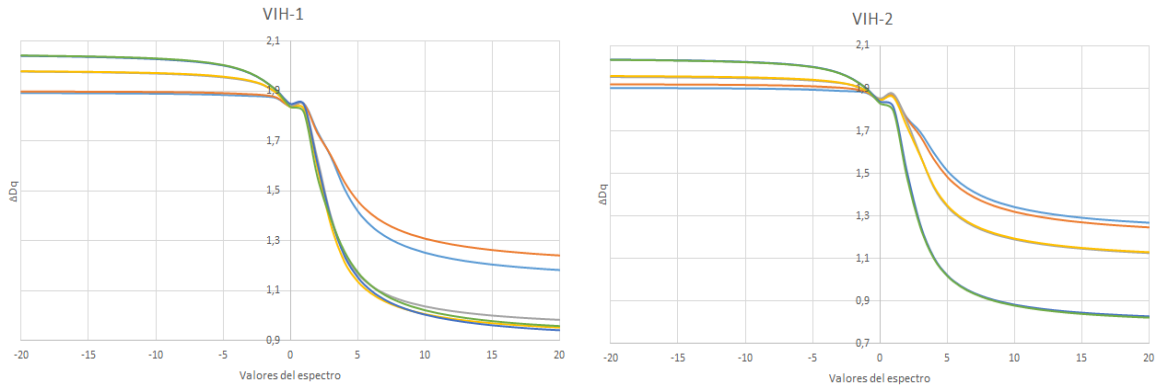

Figura 9. Espectro de distribución multifractal para el VIH-1 y el VIH-2, con el fin de evaluar el rango multifractal que son caracterizadas las secuencias de estudio.

6.3.1. Caracterización de las secuencias por valores de multifractalidad por cromosoma

Los valores de ΔDq por cromosoma, permitieron caracterizar las zonas corriente arriba y corriente a bajo que flanquean los sitios de integración del cADN lentiviral. Los valores promedio de multifractalidad obtenidos para los +100kpb corriente arriba (upstream) fueron más altos que aquellos para la ventana de -100kpb corriente abajo (downstream) (figura 10). No se encontraron diferencias significativas para los valores de multifractalidad en las dos direcciones analizada en los casos donde la homología los dos lentivirus era 100%. Para el umbral del 80-99% de homología del VIH-1, se observó una tendencia en los valores altos de multifractalidad en las zonas flanqueantes de integración corriente arriba.

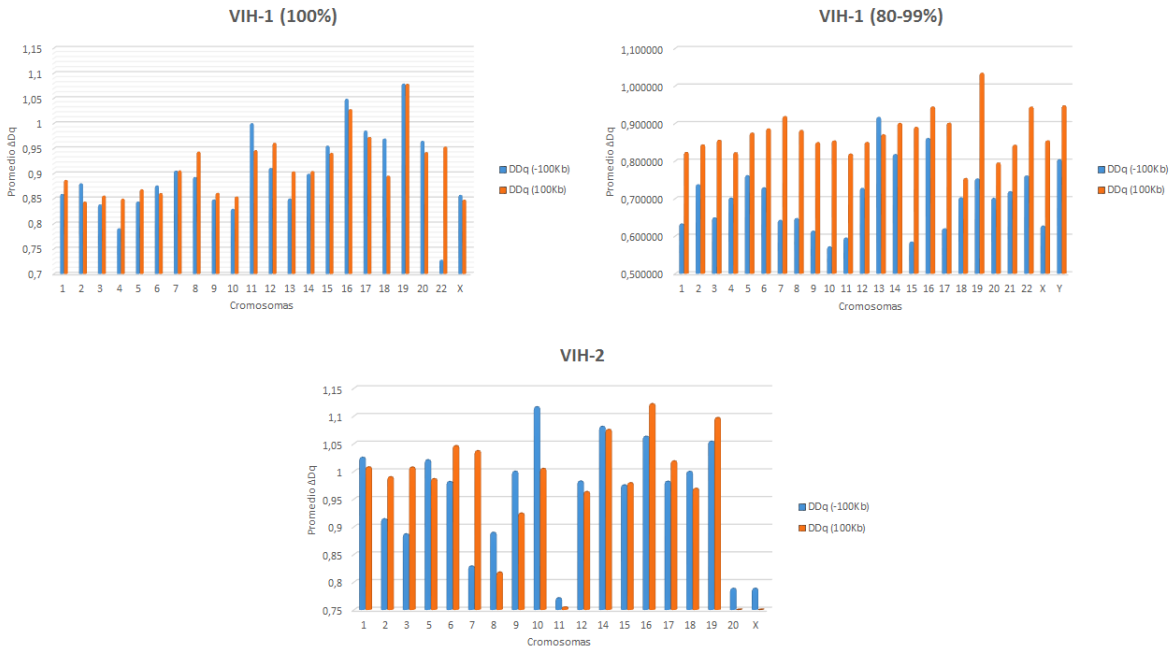

Figura 10. Distribución de los valores de los promedio de multifractalidad por secuencias y clasificado por cromosoma.

6.3.2. Rango de valores de multifractalidad de las secuencias específicas de estudio.

Puesto que los valores de multifractalidad mostraron un rango específico para aquellas secuencias de estudio de integración, se evaluó para cara tipo de provirus como era esta distribución con respecto al rango. Para el VIH-1 (100%) el 50% de las secuencias de estudio tuvieron valores entre el rango de multifractalidad de 0,85 a 0,95. Para el VIH-1 (80-99%), el 31% de las secuencias se localizó entre el rango de ΔDq de 0,75-0,9. Para el VIH-2 el 52% de las secuencias de integración proviral se incluyeron en un rango de 1,0-1,15. Estos resultados indicaron que existen diferencias entre los Lentivirus con respecto a la complejidad de las secuencias que han sido estudiadas por medio de la herramienta multifractal (Figura 11). Las secuencias adyacentes a provirus VIH-2 se encontraron en un rango con valores ΔDq mucho más altos que para las del VIH-1. Estudios previos muestran que el VIH-1 muestra una tendencia a una mayor tasa de eventos de retrotransposición

en el genoma humano que el VIH-2; además el VIH-2 tiene una menor expansión clonal. Los resultados obtenidos en este estudio, mostraron de manera indirecta que la multifractalidad es un indicador de la estabilidad integracional y de la expansión clonal en ambos Lentivirus humanos.

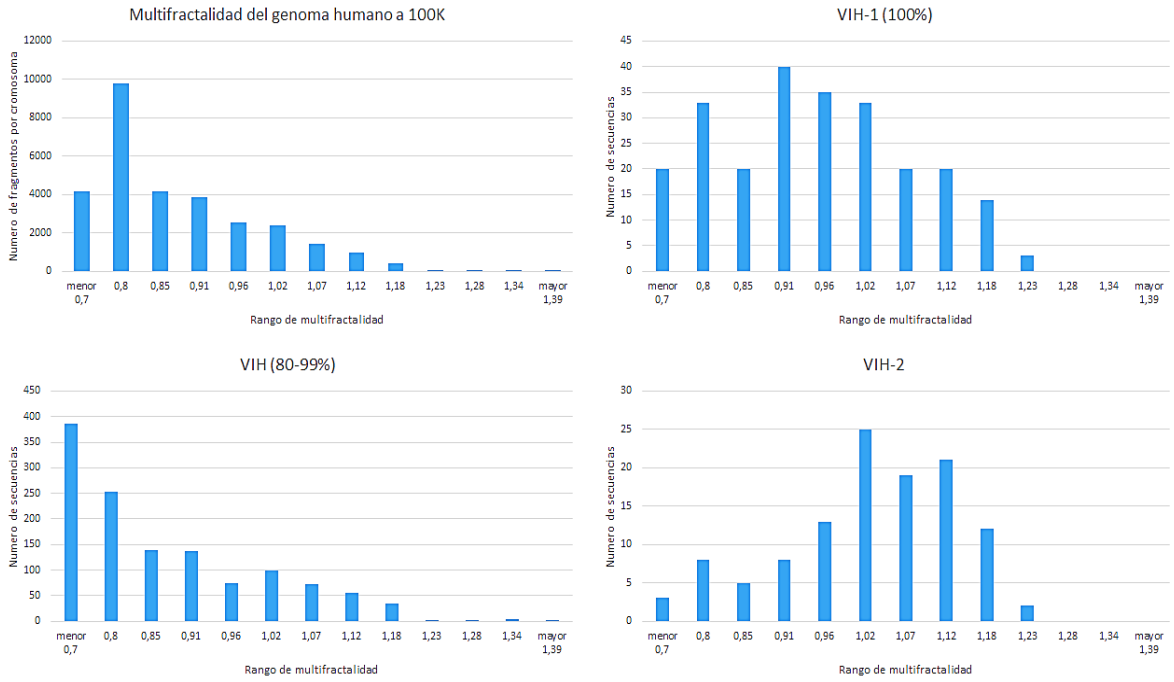

Figura 11. Distribución de las secuencias de estudio del VIH-1 y VIH-2 por rango de valores de multifractalidad.

6.3.3. Correlación entre los elementos de caracterización genómica y los valores de multifractalidad

La multifractalidad es un valor matemático que refleja la complejidad de la secuencias con respecto a sus características genómicas. Los valores de ΔDq que se obtuvieron del análisis de las secuencias flanqueantes, influyen en la preferencia de integración lentiviral del VIH-1 y VIH-2; las correlaciones de Pearson fueron muy útiles para entender como este ambiente genómico no solo es descrito por un valor matemático como el de multifractalidad, sino también por el número de estas variables que estén intrínsecamente en las secuencias incluidas en el estudio (figura

12 y 13). Tanto para el VIH-1 como para el VIH-2 se encontró una alta correlación entre las secuencias Alu, y el contenido de islas CpG; en tanto para las repeticiones LINE especialmente L1, no hubo correlación con respecto a los valores de ΔDq. Las secuencias Alu y las islas CpG son variables que contribuyen a que la cromatina sea accesible a la integración viral. Por lo general la infección por VIH de las células T por ejemplo, resulta en la integración viral con una preferencia por las regiones en el genoma humano que contienen genes activos para la expresión viral y la producción de nuevos virus, islas CpG y repeticiones como Alu. En contraposición en el estado de latencia la integración se asocia con regiones de heterocromatina constitutiva. [48].

VIH-1	Alu	DDq	L1	L2	L3	Ltr	Mer	Mir
DDq	0,851703	1	-0,0823	0,179375	-0,20791	-0,24571	-0,08979	0,05204

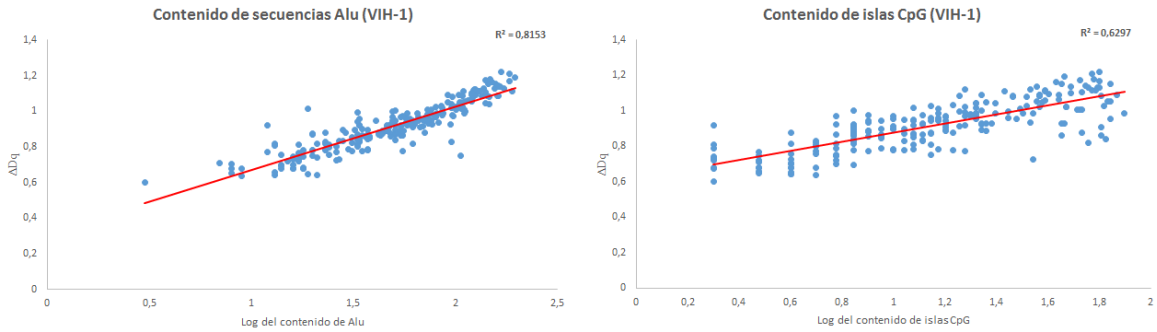

Figura 12. Valores de coeficientes de correlación de Pearson para el VIH-1 entre los valores de multifractalidad y el número de secuencias Alu e islas CpG.

VIH-2	Alu	DDq	L1	L2	L3	Ltr	Mer	Mir
DDq	0,8603	1,0000	-0,2955	-0,1354	-0,1124	-0,1626	-0,0392	-0,1348

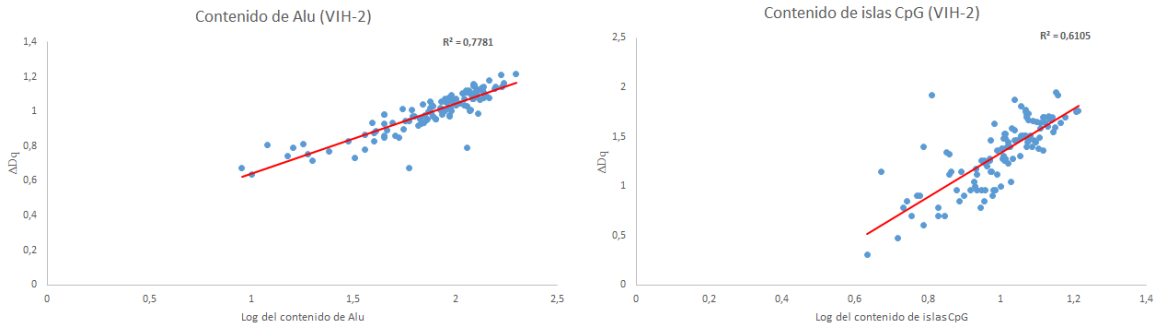

Figura 13. Valores de coeficientes de correlación de Pearson para el VIH-2 entre los valores de multifractalidad y el número de secuencias Alu e islas CpG.

El alto contenido de elementos Alu a lo largo de las secuencias, sugirió que el virus interactúa con muchas estructuras en correlación a las regiones ricas en Alu. Esto se debe que el alto contenido de Alu representa la alta aperiodicidad y la variabilidad de muchas secciones donde flanquea el provirus en el evento de la integración. Este hecho se confirmó con lo que se encontró previamente con respecto a los valores de multifractalidad que caracterizan las secuencias en general. Se observó que la integración lentiviral se encuentra en rangos de multifractalidad medios y bajos, indicando estructuralmente que la cromatina se encuentra en un estado básicamente estable y que cualquier cambio generado por el evento de integración viral, altera su estructura haciéndola más susceptible y alterando su complejidad.

Comparando los resultados obtenidos en este trabajo, con aquellos reportados para el total del genoma multifractal, se puede específicamente mencionar que la multifractalidad se relaciona principalmente con las distribuciones no lineales para las secuencias que son ricas en elementos Alu y con alto contenido de islas CpG. Los valores de multifractalidad estarían intrinsicamente relacionados con la cantidad de repeticiones Alu en las secuencias estudiadas; según el trabajo desarrollado en el genoma multifractal, entre mayor la cantidad de Alu los valores de multifractalidad tienden a ser altos puesto que estos elementos están asociados a la estabilidad y

estructura del genoma y la regulación génica en las enfermedades humanas, lo que indica que para la integración lentiviral también se encuentra implicada la gran cantidad encontrada en las secuencias de este estudio.

Un aspecto clave del estudio, fue el haber identificado y caracterizado zonas de la cromatina celular en donde ocurrió una mayor frecuencia de integración con relación a otras regiones del genoma empleando una nueva herramienta como lo es la multifractalidad; ésta nos permite estudiar la estructura de la cromatina que es susceptible a integración viral. Estas zonas de la cromatina interfásica se denominaron “hotspot” y mostraron tendencias a integración condicionadas con la localización y estado topológico de la cromatina asociada. Esta premisa se refuerza por el hecho de que las secuencias Alu, fueron la más abundante asociada a estas zonas con una correlación directa con los valores de multifractalidad obtenidos.

Es de gran importancia el estudio de las posibles funciones que son afectadas en el proceso de integración lentiviral, el agente externo ocasiona una distorsión de la información contenida por las secuencias específicas tanto en el lugar de integración y su alrededor. Los estados de la cromatina no están solamente asociados con la estructura, sino también con la función e información contenida. En este sentido se puede inferir que los sitios más polimórficos blanco de integración viral son aperiódicos; ello sugiere que las secuencias y la cromatina local están fuera del equilibrio. Sin embargo como se identificaron un mayor número secuencias con menos polimorfismos que corresponden a sitios periódicos estarían cercanos al equilibrio, y por lo tanto son más afectados por el ambiente y la remodelación de la cromatina (figura 14).

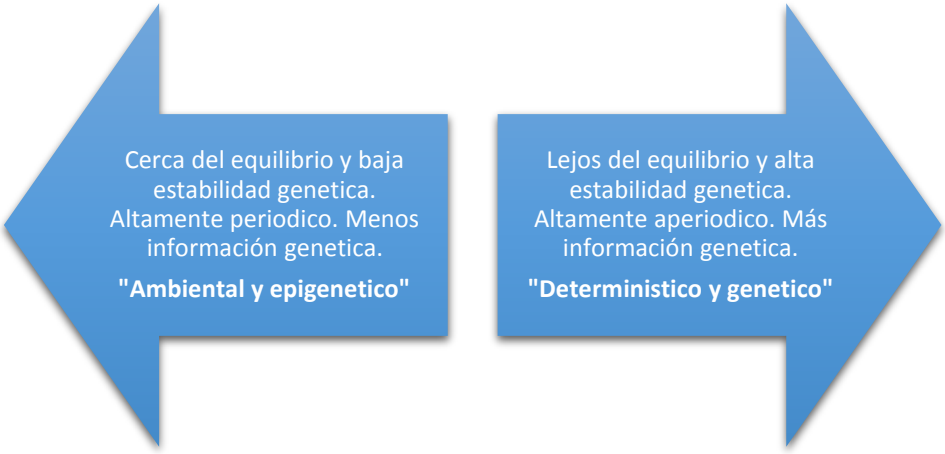

Figura 14. Diagrama que explica el significado de los valores de multifractalidad en el genoma humano y la tendencia a diferentes estados condicionados por la remodelación epigenética y ambiental y la determinística y genética.

Se determinó que los cromosomas que exhiben mayores valores de multifractalidad están asociados con territorios cromosómicos internos dentro de la cromatina interfásica [49]. La estrecha relación entre la localización de secuencias con elevada multifractalidad y los territorios cromosómicos, es evidencia muy fuerte de la existencia de una actividad transcripcional diferencial que cambia constantemente para regular múltiples procesos celulares y de activación de genes. (Figura 15).

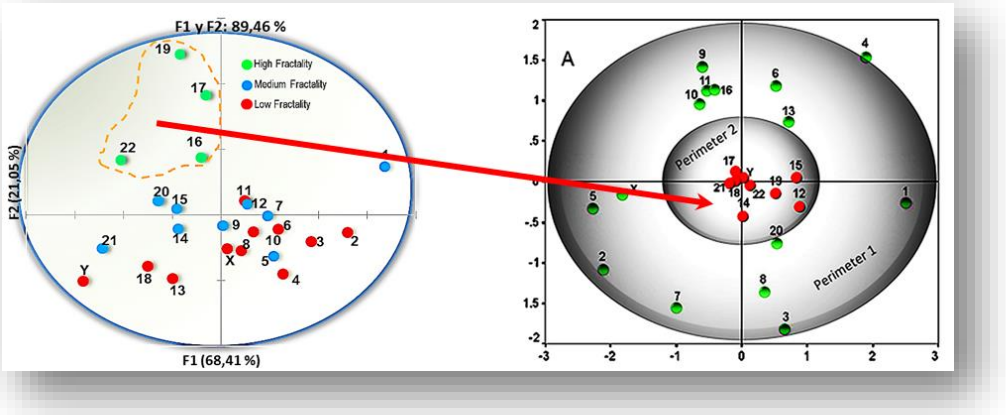

Figura 15. A) Distribución en análisis de dos componentes de los cromosomas por los valores de multifractalidad. B) Análisis de los cromosomas por territorios cromosómicos der acuerdo con resultados experimentales tomados de fibroblastos humanos en fase G0 por Bolzer et al.

La interpretación de la multifractalidad de regiones en donde el virus integra, más que un valor matemático o numérico con respecto del contenido de repeticiones e islas CpG, debe ser estudiada en función del estado de la cromatina local. Se puede inferir que zonas del genoma humano periódicas son “atractores genómicos” para las fluctuaciones ambientales dentro de las que se incluye la integración lentiviral.

6.4. Análisis de genes de las secuencias de integración de los virus VIH-1 y VIH-2

Se determinó que hay varias familias de genes localizados en las secuencias analizadas que están altamente correlacionadas con los valores de multifractalidad. Se observaron diferencias entre los genes localizados en zonas de multifractalidad variable por cromosoma para cada tipo de virus. Para el VIH-1 hubo un alto contenido de genes en las secuencias localizadas en los cromosomas 16, 17 y 19, mientras que para el VIH-2 las secuencias localizadas en los cromosomas 6 y 15 tienen mayor contenido de genes. Se ha obtenido evidencia fuerte que los cromosomas que contienen las secuencias de estudio que presentan altos valores de multifractalidad se encuentran en los cromosomas 16, 17 y 19 (figura 16). Por lo

que estos genes fueron estudiados minuciosamente para hacer seguimiento de su función el sistema inmunológico.

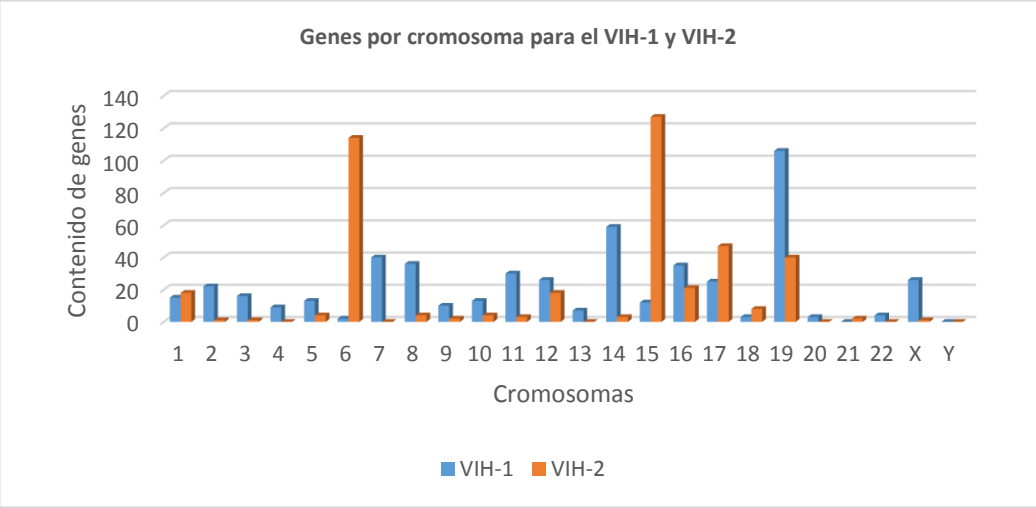

Figura 16. Contenido de genes de las secuencias del estudio por cromosoma para el VIH-1 y VIH-2.

En la cromatina existen procesos en los cuales hay disposición nucleosomica variable y dependiente de la función. Esto es el resultado de la activación ciertos promotores los cuales pueden ser más accesibles a cambios tanto génicos como ambientales, uno de estos cambios sería la integración viral que afecta áreas cuyos genes están transcripcionalmente activos por lo que estos sitios se vuelven los más frecuentes y más susceptibles al procesos de integración lentiviral.

6.5. Análisis de rutas metabólicas para los genes de las secuencias de integración de los virus VIH-1 y VIH-2

Los genes seleccionados y estudiados se localizaron en los cromosomas 16, 17 y 19 para ambos tipos de provirus; se encontró que la mayoría de estos genes pertenecen a dos rutas metabólicas esenciales: genes que codifican para factores de crecimiento que intervienen en el ciclo celular de las células T en el sistema

inmunológico y genes que son activos en la adhesión de moléculas como receptores del sistema inmunológico (figura 17).

Todos los genes se evaluaron por GO y su función específica reportada en la literatura se encuentra en las tablas anexas 1, 2 y 3.

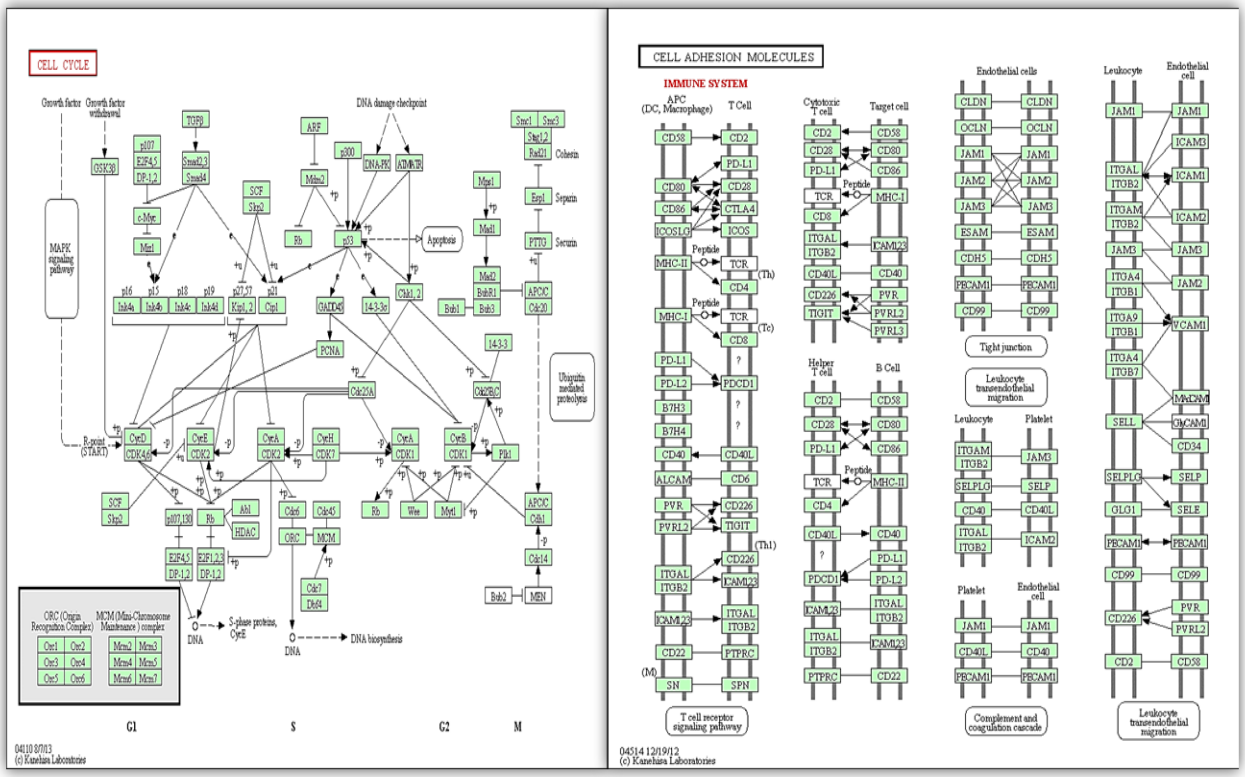

Figura 17. Rutas metabólicas predominantes en las funciones de los genes que contienen las secuencias de estudio para el VIH-1 y VIH-2 en los cromosomas 16, 17 y 19. Obtenidos de KEEG (http://www.genome.jp/kegg/).

7. CONCLUSIONES

Las estructuras caracterizadas por multifractalidad son altamente informacionales, lo que indica que el estudio por leyes de potencia no solo está caracterizando estructuralmente el fenómeno de integración lentiviral, sino que a su vez este algoritmo lo analiza de una forma más precisa. Esto es un gran avance en el estudio

de la integración lentiviral ya que las secuencias analizadas mediante la aplicación de este formalismo exhiben una alta correlación entre la multifractalidad y la frecuencia de integración lentiviral; esto confirma que los lentivirus no se integran al azar y tienen patrones específicos que pueden seguir siendo estudiados como sistemas no lineales. En general, el análisis por multifractalidad es más amplio y preciso informacionalmente, pues muestra que aquellas regiones de la cromatina celular asociadas con la integración viral están entre media y bajos valores de multifractalidad y por lo tanto corresponden a porciones inestables de la cromatina y cercanas al equilibrio.

8. PERSPECTIVAS

El conocimiento sobre el agente causal del SIDA, las relaciones entre el virus y el ser humano que lo alberga (hospedero), sus interacciones y mecanismos de daño celular, han sido fundamentales para la investigación y la terapéutica, que proporcionan una base cada vez más sólida para enfrentar y definir estrategias de control de esta epidemia [21]. El avance sistemático de este tipo investigativo acerca de otro posible factor o característica como los valores de multifractalidad de las secuencias específicas de integración lentiviral que pueden describir o definir los procesos topológicos y de complejidad que experimenta la cromatina al momento de interacción virus-hospedero nos pueden llevar a describir el ambiente genómico dinámico y la preferencia del virus a ciertas áreas definidas. Se puede considerar un gran avance científico para el avance de terapias eficaces contra la infección del SIDA.

9. BIBLIOGRAFÍA

[1] Piot P: Report on the global AIDS epidemic. UNAIDS 2011, 15, 1-367.

[2] Sharp, P. M.; Hahn, B. H. The Evolution of HIV-1 and the Origin of AIDS. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 2010, 365, 2487–2494.

[3] Vincent, K. a; York-Higgins, D.; Quiroga, M.; Brown, P. O. Host Sequences Flanking the HIV Provirus. Nucleic Acids Res. 1990, 18, 6045–6047.

[4] Mitchell, R. S.; Beitzel, B. F.; Schroder, A. R. W.; Shinn, P.; Chen, H.; Berry, C. C.; Ecker, J. R.; Bushman, F. D. Retroviral DNA Integration: ASLV, HIV, and MLV Show Distinct Target Site Preferences. PLoS Biol. 2004, 2, E234.

[5] Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; et al. Initial Sequencing and Analysis of the Human Genome. Nature 2001, 409, 860–921.

[6] Jasny, B. R.; Zahn, L. M.; Collins, F. S.; Hudson, T. A Celebration of the Genome , Part I. 2001.

[7] Hattori, M. [Finishing the Euchromatic Sequence of the Human Genome]. Tanpakushitsu Kakusan Koso. 2005, 50, 162–168.

[8] Levy, S.; Sutton, G.; Ng, P. C.; Feuk, L.; Halpern, A. L.; Walenz, B. P.; Axelrod, N.; Huang, J.; Kirkness, E. F.; Denisov, G.; Lin, Y.; MacDonald, J. R.; Pang, A. W. C.; Shago, M.; Stockwell, T. B.; Tsiamouri, A.; Bafna, V.; Bansal, V.; Kravitz, S. a; Busam, D. a; Beeson, K. Y.; McIntosh, T. C.; Remington, K. a; Abril, J. F.; Gill, J.; Borman, J.; Rogers, Y.-H.; Frazier, M. E.; Scherer, S. W.; Strausberg, R. L.; Venter, J. C. The Diploid Genome Sequence of an Individual Human. PLoS Biol. 2007, 5, e254.

[9] Dahabieh, M. S.; Ooms, M.; Brumme, C.; Taylor, J.; Harrigan, P. R.; Simon, V.; Sadowski, I. Direct Non-Productive HIV-1 Infection in a T-Cell Line Is Driven by Cellular Activation State and NFκB. Retrovirology 2014, 11, 17.

[10] Soto M J, Peña A, and García-Vallejo F: A Genomic and Bioinformatics Analysis of the Integration of HIV in Peripheral Blood Mononuclear Cells. AIDS Research and human retroviruses 2010, 549, 547-555.

[11] Moreno, P. a; Vélez, P. E.; Martínez, E.; Garreta, L. E.; Díaz, N.; Amador, S.; Tischer, I.; Gutiérrez, J. M.; Naik, A. K.; Tobar, F.; García, F. The Human Genome: a Multifractal Analysis. BMC Genomics 2011, 12, 506.

[12] Desport, M. *Lentiviruses and Macrophages: Molecular and Cellular Interactions*; Caister Academic Press, 2010.

[13] Verma, I. Lentiviral Vectors. *Nature Biotechnology*, 1999, *17*, 7–7.

[14] Garcia F: En nómada molecular, la historia del virus linfotropico humano tipo (HTLV-1). Programa editorial 2004, 44,45, 155.

[15] Hernandez D E: La infección por el virus de inmunodeficiencia humana (VIH), estudio descriptivo y experimental del compromiso de órganos y sistemas, infecciones y neoplasias. Consejo de desarrollo científico y humanístico 2002, 34, 232

[16] Fan, H.; Conner, R.; Villarreal, L. *AIDS: Science & Society*; Jones & Bartlett Learning, 2011.

[17] Soto-Girón, M. J.; García-Vallejo, F. Changes in the Topology of Gene Expression Networks by Human Immunodeficiency Virus Type 1 (HIV-1) Integration in Macrophages. Virus Res. 2012, 163, 91–97.

[18] Camargo L J, Goldberg J, Moreno F and Zulaica H: Actualidades en el tratamiento de la infección por el virus de la inmunodeficiencia humana (VIH). Anales médicos 1998, 164, 164-169.

[19] Federico M: Lentivirus Gene Engineering Protocols. Humana press 2003, 10, 299.

[20] Soto J, Peña A, Salcedo M, Domínguez M C, Sánchez A and García-Vallejo F: Genomic Characterization of HIV-1 in vitro Integration in Peripheral Blood Mononuclear Cells, Macrophages and Jurkat T Cells. Revista Infectio 2010, 22, 20-30

[21] Bushman, F. D.; Fujiwara, T.; Craigie, R. Retroviral DNA Integration Directed by HIV Integration Protein in Vitro. Science 1990, 249, 1555–1558.

[22] Crise, B.; Li, Y.; Yuan, C.; Morcock, D. R.; Whitby, D.; Munroe, D. J.; Arthur, L. O.; Wu, X. Simian Immunodeficiency Virus Integration Preference Is Similar to that of Human Immunodeficiency Virus Type 1. *J. Virol.* 2005, *79*, 12199–12204.

[23] Soto J, Peña A, Salcedo M, Domínguez M C, Sánchez A and García-Vallejo. Retroviruses, H. A Genomic and Bioinformatics Analysis of the Integration of HIV in Peripheral Blood Mononuclear Cells. 2011, 27.

[24] Rosenzweig, M.; Yamada, K.; Dm, H.; Dm, H.; T-, J. R. P.; Yamada, K.; Johnson, R. P. T-Cell Differentiation of Human and Non-Human Primate CD34 + Hematopoietic Progenitor Cells Using Porcine Thymic Stroma. 2001, 185–192.

[25] Kvaratskhelia, M.; Sharma, A.; Larue, R. C.; Serrao, E.; Engelman, A. Molecular Mechanisms of Retroviral Integration Site Selection. Nucleic Acids Res. 2014, 1–17.

[26] Mitchell, R. S.; Beitzel, B. F.; Schroder, A. R. W.; Shinn, P.; Chen, H.; Berry, C. C.; Ecker, J. R.; Bushman, F. D. Retroviral DNA Integration: ASLV, HIV, and MLV Show Distinct Target Site Preferences. PLoS Biol. 2004, 2, E234.

[27] Lewinski, M. K.; Bisgrove, D.; Shinn, P.; Chen, H.; Hannenhalli, S.; Verdin, E.; Berry, C. C.; Ecker, J. R.; Bushman, F. D.; Lewinski, M. K.; et al. Genome-Wide Analysis of Chromosomal Features Repressing Human Immunodeficiency Virus Transcription Genome-Wide Analysis of Chromosomal Features Repressing Human Immunodeficiency Virus Transcription †. 2005.

[28] Levin, A.; Armon-Omer, A.; Rosenbluh, J.; Melamed-Book, N.; Graessmann, A.; Waigmann, E.; Loyter, A. Inhibition of HIV-1 Integrase Nuclear Import and Replication by a Peptide Bearing Integrase Putative Nuclear Localization Signal. Retrovirology 2009, 6, 112.

[29] Nishimura, Y., Sadjadpour, R., Mattapallil, J. J., Igarashi, T., Lee, W., Buckler-White, A., Roederer, M., Chun, T.-W., and Martin, M. a. (2009) High frequencies of resting CD4+ T cells containing integrated viral DNA are found in rhesus macaques during acute lentivirus infections. Proc. Natl. Acad. Sci. U. S. A. 106, 8015–20.

[30] Muller W: Bioquímica, fundamentos para medicina y ciencias de la vida. Editorial Reverte 2008, 333, 560.

[31] Anonymous. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science. 1998, 282, 2012-2018.

[32] Berthelsen CL, Glazier JA and Raghavachari S (1994). Effective multifractal spectrum of a random walk. Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics 49, 1860-1864.

[33] Vélez, P. E.; Garreta, L. E.; Martínez, E.; Díaz, N.; Amador, S.; Tischer, I.; Gutiérrez, J. M.; Moreno, P. a. The Caenorhabditis Elegans Genome: A Multifractal Analysis. Genet. Mol. Res. 2010, 9, 949–965.

[34] Burgos JD and Moreno-Tovar P. Zipf-scaling behavior in the immune system. Biosystems. 1996, 39, 227-232.

[35] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A,

Sougnez C, et al: Initial sequencing and analysis of the human genome. Nature 2001, 409, 860-921.

[36] Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, et al: The diploid genome sequence of an individual human. PLoS Biol 2007, 5, e254.

[37] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor GLMiklos, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, et al: The sequence of the human genome. Science 2001, 291, 1304-51.

[38] Versteeg R, van Schaik BDC, van Batenburg MF, Roos M, Monajemi R, Caron H, Bussemaker HJ, van Kampen AHC: The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. Genome Research 2003, 13, 1998-2004.

[39] DeCerbo J, Carmichael GG: SINEs point to abundant editing in the human genome. Genome Biology 2005, 6:216. 7. The 1000 genomes project consortium: A map of human genome variation from population-scale sequencing. Nature 2010, 467:1061-1073.

[40] Moreno PA, Vélez PE, Burgos JD: Biología molecular, genómica y postgenómica. Pioneros, principios y tecnologías Editorial Universidad del Cauca: Popayán; 2009.

[41] Restrepo S, Pinzón A, Rodríguez-R LM, Sierra R, Grajales A, Bernal A, Barreto E, Moreno PA, Zambrano MM, Cristancho M, González A, Castro H: Computational biology in Colombia. PLoS Comput Biol 2009, 5(10), e1000535.

[42] Mandelbrot B: La geometría fractal de la naturaleza Tusquets editores: Barcelona; 1982.

[43] Clark, A. G.; Hubisz, M. J.; Bustamante, C. D.; Williamson, S. H.; Nielsen, R. Ascertainment Bias in Studies of Human Genome-Wide Polymorphism. *Genome Res.* 2005, *15*, 1496–1502.

[44] Stevens, S. W., and J. D. Griffith. 1994. Human immunodeficiency virus type 1 may preferentially integrate into chromatin occupied by L1Hs repetitive elements. Proc. Natl. Acad. Sci. USA 91, 5557–5561.

[45] Stevens, S. W., and J. D. Griffith. 1996. Sequence analysis of the human DNA flanking sites of human immunodeficiency virus type 1 integration. J. Virol. 70, 6459–6462.

[46] Zimbwa P, Milicic A, Frater J, Scriba TJ, Willis A, Goulder PJR, et al. Precise identification of a human immunodeficiency virus type 1 antigen processing mutant. J Virol. 2007; 81 (4), 20331–2038.

[48] Gallastegui, E.; Millán-Zambrano, G.; Terme, J.-M.; Chávez, S.; Jordan, A. Chromatin Reassembly Factors Are Involved in Transcriptional Interference Promoting HIV Latency. J. Virol. 2011, 85, 3187–3202.

[49] Gagniuc, P.; Ionescu-Tirgoviste, C. Gene Promoters Show Chromosome-Specificity and Reveal Chromosome Territories in Humans. BMC Genomics 2013, 14, 278.

10. ANEXOS

Genes Crm. 16	ID	Nombre
MRPL28	10573	mitochondrial ribosomal protein L28
TMEM8A	58986	transmembrane protein 8A
RPL23AP5	729480	ribosomal protein L23a pseudogene 5
LOC100996484	100996484	uncharacterized LOC100996484
NME4	4833	NME/NM23 nucleoside diphosphate kinase 4
DECR2	26063	2,4-dienoyl CoA reductase 2, peroxisomal
HN1L	90861	hematological and neurological expressed 1-like
LOC101929480	101929480	uncharacterized LOC101929480
MIR3177	100423012	microRNA 3177
NME3	4832	NME/NM23 nucleoside diphosphate kinase 3
MRPS34	65993	mitochondrial ribosomal protein S34
EME2	197342	essential meiotic structure-specific endonuclease subunit 2
SPSB3	90864	splA/ryanodine receptor domain and SOCS box containing 3
NUBP2	10101	nucleotide binding protein 2
IGFALS	3483	insulin-like growth factor binding protein, acid labile subunit
NLRC3	197358	NLR family, CARD domain containing 3
LOC101929732	101929732	uncharacterized LOC101929732
SLX4	84464	SLX4 structure-specific endonuclease subunit
TRAP1	10131	TNF receptor-associated protein 1
LOC101929751	101929751	uncharacterized LOC101929751
PLA2G10	8399	phospholipase A2, group X
LOC100652777	100652777	uncharacterized LOC100652777
NPIPA3	642778	nuclear pore complex interacting protein family, member A3
IL17C	27189	interleukin 17C
CYBA	1535	cytochrome b-245, alpha polypeptide
MVD	4597	mevalonate (diphospho) decarboxylase
SNAI3-AS1	197187	SNAI3 antisense RNA 1
SNAI3	333929	snail family zinc finger 3
RNF166	115992	ring finger protein 166
CTU2	348180	cytosolic thiouridylase subunit 2 homolog (S. pombe)
MIR4722	100616167	microRNA 4722
LOC100289580	100289580	uncharacterized LOC100289580
CDT1	81620	chromatin licensing and DNA replication factor 1
APRT	353	adenine phosphoribosyltransferase

Tabla anexo 1. Genes localizados en las secuencias de integración del VIH del cromosoma 16.

Genes Crm. 17	ID	Nombre
TM4SF5	9032	transmembrane 4 L six family member 5
VMO1	284013	vitelline membrane outer layer 1 homolog (chicken)
GLTPD2	388323	glycolipid transfer protein domain containing 2
PSMB6	5694	proteasome (prosome, macropain) subunit, beta type, 6
PLD2	5338	phospholipase D2
ATP6V0CP1	100132978	ATPase, H+ transporting, lysosomal 16kDa, V0 subunit c pseudogene 1
CHRNE	1145	cholinergic receptor, nicotinic, epsilon (muscle)
C17orf107	100130311	chromosome 17 open reading frame 107
GP1BA	2811	glycoprotein Ib (platelet), alpha polypeptide
SLC25A11	8402	solute carrier family 25 (mitochondrial carrier; oxoglutarate carrier), member 11
RNF167	26001	ring finger protein 167
PFN1	5216	profilin 1
MIR4728	100616132	microRNA 4728
MIEN1	84299	migration and invasion enhancer 1
GRB7	2886	growth factor receptor-bound protein 7
KRT8P34	100418811	keratin 8 pseudogene 34
RPL39P4	654392	ribosomal protein L39 pseudogene 4
ZPBP2	124626	zona pellucida binding protein 2
GSDMB	55876	gasdermin B
CEP131	22994	centrosomal protein 131kDa
ENTHD2	146705	ENTH domain containing 2
C17orf89	284184	chromosome 17 open reading frame 89
LINC00482	284185	long intergenic non-protein coding RNA 482
TMEM105	284186	transmembrane protein 105

Tabla anexo 2. Genes localizados en las secuencias de integración del VIH del cromosoma 17.

Genes Crm. 19	ID	Nombre
SMIM24	284422	small integral membrane protein 24
DOHH	83475	deoxyhypusine hydroxylase/monooxygenase
MFSD12	126321	major facilitator superfamily domain containing 12
C19orf71	100128569	chromosome 19 open reading frame 71
HMG20B	10362	high mobility group 20B
GIPC3	126326	GIPC PDZ domain containing family, member 3
TBXA2R	6915	thromboxane A2 receptor
CACTIN-AS1	404665	CACTIN antisense RNA 1
CREB3L3	84699	cAMP responsive element binding protein 3-like 3
SIRT6	51548	sirtuin 6
EBI3	10148	Epstein-Barr virus induced 3
CCDC94	55702	coiled-coil domain containing 94
SHD	56961	Src homology 2 domain containing transforming protein D
TMIGD2	126259	transmembrane and immunoglobulin domain containing 2
FSD1	79187	fibronectin type III and SPRY domain containing 1
STAP2	55620	signal transducing adaptor family member 2
MPND	84954	MPN domain containing
CHAF1A	10036	chromatin assembly factor 1, subunit A (p150)
UBXN6	80700	UBX domain protein 6
MIR4746	100616371	microRNA 4746
C19orf66	55337	chromosome 19 open reading frame 66
ANGPTL6	83854	angiopoietin-like 6
PPAN-P2RY11	692312	PPAN-P2RY11 readthrough
PPAN	56342	peter pan homolog (Drosophila)
SNORD105	692229	small nucleolar RNA, C/D box 105
SNORD105B	100113382	small nucleolar RNA, C/D box 105B
P2RY11	5032	purinergic receptor P2Y, G-protein coupled, 11
EIF3G	8666	eukaryotic translation initiation factor 3, subunit G
S1PR2	9294	sphingosine-1-phosphate receptor 2
MIR4322	100422925	microRNA 4322
MRPL4	51073	mitochondrial ribosomal protein L4
ICAM4	3386	intercellular adhesion molecule 4 (Landsteiner-Wiener blood group)
ICAM5	7087	intercellular adhesion molecule 5, telencephalin
ZGLP1	100125288	zinc finger, GATA-like protein 1
FDX1L	112812	ferredoxin 1-like
RAVER1	125950	ribonucleoprotein, PTB-binding 1
ICAM3	3385	intercellular adhesion molecule 3
CDC37	11140	cell division cycle 37
MIR1181	100302213	microRNA 1181
PDE4A	5141	phosphodiesterase 4A, cAMP-specific
CD97	976	CD97 molecule

Tabla anexo 3. Genes localizados en las secuencias de integración del VIH del cromosoma 19.